UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Matematica “Tullio-Levi Civita”

UNIVERSITÉ PARIS-DAUPHINE PSL

Département de Mathématiques et Applications

Master’s Degree Course in Mathematics

Double Degree Course “MAPPA”

KERNEL STATISTICAL DISTANCES AND APPLICATION TO NLP

Advisor:
Prof. Gabriel Turinici
Co-Advisor:
Prof. Daniela Tonon
MAPPA project coordinator:
Prof. Francesco Rossi

Student:
Cecilia Secchi

Student ID:
2021084 / 22100603

Academic year 2021/2022

Contents

Introduction

The purpose of this master's thesis is to apply different types of statistical distances to natural language processing. In particular, it seeks to find keywords from given sentences, where keywords are defined as words of particular relevance that can help to understand the meaning and content of the sentence.
The problem of automatic analysis and summarization of the content of written sentences has become of great importance with the rise of the online world and the huge amount of information created daily by users and customers to provide their opinions and feedback in e-commerce and social networks. For example, companies may be interested in studying this written content to decide how to manage their organization and future investments, as well as public figures may learn about followers' reactions to their posts.

To work with written content, the first thing to do is to turn words into vectors. For this purpose we used GloVe embedding, a dictionary created by Stanford researchers that associates semantically close words with geometrically close vectors. In this way we translated our sentences into an ordered list of vectors, where each vector corresponds to a word.

Our original approach to find the keyword consists of this: we interpret the sentence as a probability distribution and, using statistical distances, try to compress it into a new probability distribution of the form of a small sum of Diracs. The vectors supporting this distribution will correspond to the keywords we are looking for.
To calculate the best approximate distribution we use Kernel statistical distances, which are a generalization of the energy distance. The Energy Distance between μ and ν is defined as

$$\mathcal{E}(\mu,\nu) = 2\mathbb{E}[\|X-Y\|] - \mathbb{E}[\|X-X'\|] - \mathbb{E}[\|Y-Y'\|]$$

where X, X', Y, Y' are independent random variables, with law $X, X' \sim \mu$ and $Y, Y' \sim \nu$.

Statistical kernel distances replace the Euclidean norm present in the expected values with some kernels that must meet certain properties. The use of different kernel functions allows us to create versatile statistical distances that can best fit the specific problem to be solved.

The first chapter contains a detailed explanation of how GloVe incorporation works. Then there is also a description of how tf-idf works. Tf-idf is a common NLP weighting method that values words of significant meaning.

The second chapter defines energy distance and contains evidence that it respects the properties of distance. Then there is a characterization of a class of kernels that can be used to create new statistical distances. This is followed by a description of the optimization algorithm that allows us to find the keywords, given the sentence.

The third chapter contains the experiments performed and the results obtained. In particular, in addition to kernel statistical distances, it also contains other approaches such as k-mean clustering.

Chapter 1

Natural Language Process

In this chapter we will explain the method we applied to transform the words into vectors.
We used GloVe embedding, a model implemented using as input a co-occurrence matrix that, given a large corpus, records the frequency with which two words occur in the same sentence. Taking advantage of this relationship, the model produces vectors that take into account the semantic closeness of words: words that are semantically close will be associated with neighboring vectors.

1.1 GloVe

GloVe is an unsupervised learning algorithm for obtaining vector representations for words [8]. These vectors are constructed in such a way that neighboring vectors correspond to semantically close words.
With respect to the other fundamental word embedding *Word2Vec*, see [6], GloVe doesn't rely only on local statistics (local context information of the words), but incorporate global statistics to obtain word vectors.

The fundamental object of the GloVe model is the matrix that collects global word-word co-occurrence statistics from a large corpus. The model is trained to minimize a specific loss function in which the relationship between words is given by the matrix.

The model produces a word-vector space with meaningful linear substructure. Classical examples that express this relation are:

- king - man + woman = queen

- paris - france + germany = berlin

The difference vector *paris - france* captures the concept of *capital city*, the difference vector *king - man* captures the concept of gender-high power.

We used the Glove database available in the Stanford website of 6B tokens and that returns vectors of dimension 50.

1.1.1 Implementation of GloVe

In this section we will see how the model finds the vectors associated with each word. It will be necessary to choose a specific weighted least-squares function that contains the global co-occurrence count information, and minimizing that function among the embedded words will produce the desired result.

We denote with letters such as i, j the words of our training corpus. X is the matrix of co-occurrence, where X_{ij} represents the number of times word j occurs in the context of word i. $X_i = \sum_k X_{ik}$ is the number of times any word appears in the context of word i.
Finally we call $P_{ij} = P(j|i) = X_{ij}/X_i$ the probability of find word j in the context of word i.

It has been chosen to consider the ratio of the co-occurrence probabilities rather than the raw ones as it seems to give more information: the data P_{ik}/P_{jk} is better able to distinguish relevant words k for i or j with respect to irrelevant ones:

- if the word k is relevant for i and not for j, then the ratio will be bigger than 1;
- if the word k is relevant for j and not for i, then the ratio will be smaller than 1;
- if the word k is correlated or not correlated with both i and j, then the ration will be close to 1.

After this preliminary considerations, we look for a function F that associates the probability ratio to each triplet of vectors corresponding to i, j, k:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \tag{1.1}$$

Probability and Ratio	k = solid	k = gas	k = water	k = fashion
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Table 1.1: Example of probabilities from X: *solid* is relevant just for *ice*, *gas* is relevant just for *steam*, *water* is relevant for both, *fashion* for neither. Source:[8]

where $w \in \mathbb{R}^d$ are the target word vectors and $\tilde{w} \in \mathbb{R}^d$ are separate context word vectors.
It has been chosen

$$F(w_i, w_j, \tilde{w}_k) = \exp((w_i - w_j)^T \tilde{w}_k) = \frac{\exp(w_i^T \tilde{w}_k)}{\exp(w_j^T \tilde{w}_k)} \tag{1.2}$$

as it satisfies the following properties:

- F depends on the difference of the target words $w_i - w_j$ as we want the information present in the ratio P_{ik}/P_{jk} to be encoded in a vector space;
- the vectors of F have been transformed to a scalar through the dot product, in order to have the same type of the ratio;
- F is a omomorphism between $(\mathbb{R}, +)$ and $(\mathbb{R}_{>0}, \cdot)$, and in this way is symmetric with respect to $w_i - w_j$ and $\tilde{w}_k$.

So observing equations 1.1 and 1.2, , up to multiplication for a constant, we want the vectors $w, \tilde{w}$ to satisfy:

$$\exp(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}. \tag{1.3}$$

and hence also:

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i). \tag{1.4}$$

The term $\log(X_i)$ disturbs the symmetry of the equation, so we add to the expression the bias terms $b, \tilde{b}$ per $w, \tilde{w}$ and we make them absorb X_i, obtaining:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}). \tag{1.5}$$

At this point we add a weighted function f in such a way that the log will not explode when X_{ik} will be 0 or in its neighborhood because of the logarithm. Moreover we also want f not to give a lot of importance to rare co-occurrences, so we choose a function f with the following properties:

1. f is continuous, $f(0) = 0$ and $\lim_{x\to 0} f(x)\log^2 x$ is finite;
2. f is increasing, so that rare co-occurrences have small weight;
3. f is bounded from above, so that frequent co-occurrences are not overweighted.

For the implementation of the model, it has been chosen the family of functions:

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \tag{1.6}$$

with $\alpha = 3/4$,$x_{\max} = 100$, found empirically.
So in conclusion our model finds the word-vectors minimizing the cost function

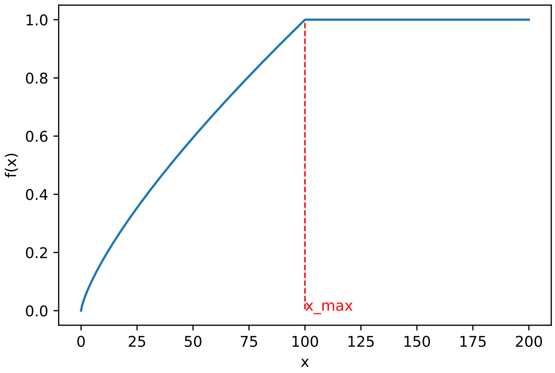

Figure 1.1: Function f with parameters $\alpha = 3/4, x_{\max} = 100$.

J:

$$J(w, \tilde{w}, b, \tilde{b}) = \sum_{i,j} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{1.7}$$

The model returns for each word k the vectors w_k and $\tilde{w}_k$, as it consider it both as a context word and as a context one. The average of the two vectors is returned as the word vector.

Training of the model

We call the context window for a given word the width of the text in which to search for co-occurrences. The model is trained using symmetrical context windows of 10 words on the left and 10 words on the right.

The minimization of the loss function J is done using the Adaptive Gradient Descent algorithm (AdaGrad) [3],a modified Stochastic Gradient Descent. This algorithm stochastically samples non zero elements from X, with initial learning rate of $\alpha = 0$, and at each iteration it updates the learning rate for each parameter independently of the others, incorporating knowledge of past observations. It performs larger updates (e.g. high learning rates) for those parameters that are related to infrequent features and smaller updates (i.e. low learning rates) for frequent one. As a result, it is well-suited when dealing with sparse data, as in our case.

The number of iterations is related to the dimension of the vector (50 iterations for word vectors of dimension less than 300, 100 otherwise).

Observations:

- The worst case scenario of the complexity of the model is $\mathcal{O}(\|V\|^2)$ where $\|V\|$ is the size of the vocabulary.
- The model reaches a 75% accuracy performance in the word analogy task proposed, that consists of questions like: a is to b as c is to __? For example "Athens is to Greece as Berlin is to?" The word d that answers the question is the one whose vector representation w_d is the closest to $w_b - w_a + w_c$ according to the cosine similarity.

1.2 Tf-Idf

In information retrieval, tf-idf, short for term frequency-inverse document frequency, is a numerical statistic that reflects the importance of a word t for a document d in a collection of documents D [18]. tf-idf is one of the most popular term weighting schemes today. The value of tf-idf increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. This helps to adjust for the fact that some words appear more frequently in general.

The weight is composed by two factors: TF and IDF.

- **tf**: Term Frequency, it measures how frequently a term occurs in a document. Since every document is different in length, the term frequency is often divided by the total number of terms in the document as a way of normalization:

$$\mathrm{TF}(t,d) = \frac{\text{Number of times term t appears in a document d}}{\text{Total number of terms in the document d}}.$$

In this way a term that belongs to long documents and then that appear much more times than a term belonging to shorter ones is not unfairly weighted.

- **IDF**: Inverse Document Frequency, it measures how important a term is. Words such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$\mathrm{idf}(t,D) = \frac{\log(\text{Total number of documents } D)}{\text{Number of documents with term } t \text{ in it}}.$$

- **tf-idf**: product of tf and idf:

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

  A high weight for a word t is reached if t has high frequency in the given document d and a low document frequency in the whole collection of documents; the weights hence tend to filter out common terms.

Since the ratio inside the idf's log function is always greater than or equal to 1, the value of idf (and tf–idf) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the idf and tf–idf closer to 0.

Chapter 2

Theory of the distances

In this chapter we will study the theory behind the distances we will use in the experimental part. We will discuss kernel distances, a family of statistical distances derived from the energy distance.
We will first give the definition of energy distance and the property that a kernel must have to preserve the non-negativity of the distance function. We will show a nice proof of the non-negativity of the energy distance. We will then study a family of kernel distances in which the energy distance is included. Characterization of the associated distances through the characteristic functions of the argument distributions will allow us to conclude that the three metric properties are satisfied for these statistical kernel distances.
We will conclude the chapter by discussing geodesic kernels and k-mean clustering.

2.1 Energy distance

Energy statistics are functions of distances between statistical observations introduced by Székely and Rizzo [13]. This concept is based on the notion of Newton's gravitational potential energy which is a function of the distance between two bodies.

Remark 1. *In the following, unless specified otherwise, X, Y will always be independent random variables defined on $\mathbb{R}$ or $\mathbb{R}^d$, with $\mathbb{E}[\|X\|] < \infty$, $\mathbb{E}[\|Y\|] < \infty$. X', Y' will be other independent random variables with the same distribution of respectively X, Y.*

Definition 1. *The Energy Distance between the d-dimensional independent* *random variables X and Y is defined as*

$$\mathcal{E}(X,Y) = 2\mathbb{E}[\|X-Y\|] - \mathbb{E}[\|X-X'\|] - \mathbb{E}[\|Y-Y'\|] \tag{2.1}$$

This distance has the property to be rotation invariant. In the 1-dimensional case the distance is equal to the double of the Cramér-von Mises distance [2], as it is proved in the following theorem [14].

Theorem 1. *If X, Y are independent random variables on $\mathbb{R}$ with cumulative distribution functions F, G, then*

$$\mathcal{E}(X,Y) = 2\int_{\mathbb{R}} (F(t)-G(t))^2 dt \tag{2.2}$$

Proof. Let us begin rewriting in an other form the right hand side:

$$\begin{aligned} 2\int_{\mathbb{R}} &(F(t)-G(t))^2 dt \\ &= 2\int_{\mathbb{R}} F(t)(1-G(t)) + (1-F(t))G(t) \\ &\quad - F(t)(1-F(t)) - G(t)(1-G(t))dt \\ &= 2\int_{\mathbb{R}} \mathbb{P}(X \le t < Y) + \mathbb{P}(Y \le t < X) \\ &\qquad - \mathbb{P}(X \le t < X') - \mathbb{P}(Y \le t < Y')dt \end{aligned}$$

where in the last line we have used the independence of X, X', Y, Y' and hence the equality

$$\mathbb{P}(X \le t) \cdot \mathbb{P}(Y > t) = \mathbb{P}(X \le t < Y). \tag{2.3}$$

To conclude the proof, it is enough to show that

$$\mathbb{E}[|X-Y|] = \int_{\mathbb{R}} \mathbb{P}(X \le t < Y) + \mathbb{P}(Y \le t < X)dt. \tag{2.4}$$

Let first observe that

$$\mathbb{E}[|X-Y|] = \mathbb{E}[(X-Y)\mathbb{1}_{X>Y}] + \mathbb{E}[(Y-X)\mathbb{1}_{Y>X}]$$

then we use Fubini's theorem and the independence of X, Y:

$$\begin{aligned} \mathbb{E}[|X-Y|] &= \int_{\Omega}\int_{\mathbb{R}} \mathbb{1}_{Y(\omega)\le t<X(\omega)} + \mathbb{1}_{Y(\omega)\le t<X(\omega)} dt d\omega \\ &= \int_{\mathbb{R}}\int_{\Omega} \mathbb{1}_{Y(\omega)\le t<X(\omega)} + \mathbb{1}_{Y(\omega)\le t<X(\omega)} d\omega dt \\ &= \int_{\mathbb{R}} \mathbb{P}(X \le t < Y) + \mathbb{P}(Y \le t < X)dt. \end{aligned}$$

If X, Y have the same distribution, then $\mathbb{P}(X \le t < Y) = \mathbb{P}(Y \le t < X)$ and this ends the proof. □

2.1.1 Kernel Statistical Distances

The Energy Distance can be generalized using different functions (kernels) instead of the usual Euclidean norm, as long as the positivity of the expression 2.1 is guaranteed.
The following will provide a characterization of functions for which the generalized energy distance is non-negative. These functions are called the *negative definite kernels.*
We will use this characterization to prove that for the euclidean norm and any power α of it for $\alpha \in (0, 2)$, the energy distance is non-negative.

Definition 2. *The Kernel Distance for the kernel $\mathcal{L} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is of the form:*

$$\mathcal{E}(\mathcal{L}(X, Y)) = 2\mathbb{E}[\mathcal{L}(X, Y)] - \mathbb{E}[\mathcal{L}(X, X')] - \mathbb{E}[\mathcal{L}(Y, Y')], \tag{2.5}$$

provided that $\mathcal{L}$ preserves the distance properties.

Definition 3. *Let $\mathcal{X}$ be a set and $\mathcal{L}(x, y)$ a hermitian function defined on $\mathcal{X} \times \mathcal{X}$ and assuming complex values.*
We say that $\mathcal{L}(x, y)$ is a negative definite kernel if for any integer $n \geq 1$, any point $x_1, \ldots, x_n \in \mathcal{X}$, and any complex number $h_1 \ldots, h_n$, the inequality

$$\sum_{s=1}^{n} \sum_{t=1}^{n} \mathcal{L}(x_s, x_t) h_s \bar{h_t} \leq 0 \tag{2.6}$$

holds under the condition $\sum_{j=1}^{n} h_j = 0$.
The function $\mathcal{L}(x, y)$ is called strictly negative-definite if equality is attained in 2.6 only when $h_j = 0$ for all $j = 1, \ldots, n$.

If the function $\mathcal{L}(x, y)$ assumes only real values and it is symmetric, that is for all $x, y \in \mathcal{X}$

$$\mathcal{L}(x, y) = \mathcal{L}(y, x), \tag{2.7}$$

then it suffices to take only real h_j in 2.6. For, if $h_i = a_i + ib_i$, with $a_i, b_i \in \mathbb{R}$, then

$$\sum_{i,j=1}^{n} \mathcal{L}(x_i, x_j) h_i \bar{h_j} = \sum_{i,j=1}^{n} \mathcal{L}(x_i, x_j)(a_i a_j + b_i b_j) + i \sum_{i,j=1}^{n} \mathcal{L}(x_i, x_j)(b_i a_j - b_j a_i) \tag{2.8}$$

and the last sum is 0 if $\mathcal{L}$ is symmetric.

Now we assume $\mathcal{X}$ to be a metric space with distance d. We denote by $\mathcal{B}$ the σ-algebra of Borel subsets of $\mathcal{X}$. We will study for which kind of kernel $\mathcal{L}$ the Kernel Distance preserves the distance property:

$$\mathcal{E}(\mathcal{L}(X, Y)) \geq 0 \tag{2.9}$$

Theorem 2. *Let $\mathcal{L}(x,y)$ be a real-valued continuous function satisfying condition 2.7. Inequality 2.9 holds for all mutually independent random variables X, X', Y, Y' with values in $(\mathcal{X}, \mathcal{B})$ satisfying the conditions*

$$\mathbb{E}[\mathcal{L}(X,X')] < \infty, \mathbb{E}[\mathcal{L}(Y,Y')] < \infty \tag{2.10}$$

if and only if the function $\mathcal{L}$ is a negative definite kernel.

Proof. From [19]. The definition of a symmetric negative-definite kernel can be rewritten in an equivalent form as

$$\iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)h(x)h(y)dQ(x)dQ(y) \leq 0 \tag{2.11}$$

for any probability measure Q on $(\mathcal{X}, \mathcal{B})$ and any integrable function $h(x)$ with $\int_{\mathcal{X}} h(x)dQ = 0$.
We introduce measures μ and ν on $\mathcal{B}$ by setting for all $A \in \mathcal{B}$:

$$\mu(A) = \mathbb{P}\{X \in A\}, \quad \nu(A) = \mathbb{P}\{Y \in A\}. \tag{2.12}$$

Let Q_1 be an arbitrary probability on $(\mathcal{X}, \mathcal{B})$ dominating μ and ν, so such that $\mu \ll Q_1$ and $\nu \ll Q_1$. Hence there are two functions h_1 and h_2 for which we can write

$$d\mu = h_1(x)dQ_1, \quad d\nu = h_2(x)dQ_1. \tag{2.13}$$

Defining $h(x) := h_1(x) - h_2(x)$, we can write:

$$\begin{aligned}
2\mathbb{E}[\mathcal{L}(X,Y)] - \mathbb{E}[\mathcal{L}(X,X')] - \mathbb{E}[\mathcal{L}(Y,Y')] = \\
\iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)d\mu(x)d\nu(y) + \iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)d\mu(y)d\nu(x) \\
- \iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)d\mu(x)d\mu(y) - \iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)d\nu(x)d\nu(y) \\
= - \iint_{\mathcal{X}\times\mathcal{X}} \mathcal{L}(x,y)h(x)h(y)dQ_1(x)dQ_1(y) \geq 0
\end{aligned} \tag{2.14}$$

and

$$\int h(x)dQ_1(x) = 0. \tag{2.15}$$

And thanks to the arbitrariness of Q_1 and h, we can conclude.

□

Now we will use this Theorem to prove that the Energy Distance of Definition 1, that is the kernel distance where the kernel is the euclidean norm, is non-negative.

Proposition 1. *Let $\mathcal{L}(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be defined as the Euclidean distance between vectors x, y. Then $\mathcal{L}(x, y) = \|x - y\|$ is strictly negative definite.*

Proof. [12] Let $x_1, \dots, x_n$ be arbitrary distinct point in $\mathbb{R}^d$ and $r_1, \dots, r_n$ be real numbers. The thesis is that:

$$\sum_{i,j=1}^{n} \|x_i - x_j\| r_i r_j \leq 0 \quad \text{if} \quad \sum_{i=1}^{n} r_i = 0 \tag{2.16}$$

with equality if and only if $r_1 = \dots = r_n = 0$.
We can express 2.16 in an equivalent form, observing that each r_i can be rewritten as $r_i = p_i - q_i$ for $p_i, q_i \in \mathbb{R}_+$ and that $\sum_i r_i = 0$ implies $\sum_i p_i = \sum_i q_i$. So the thesis becomes:
if $x_1, \dots, x_n$ are distinct arbitrary points in $\mathbb{R}^d$ and $p_1, \dots, p_n$ and $q_1, \dots, q_n$ are probability distributions defined on them, then

$$\sum_{i,j=1}^{n} \|x_i - x_j\| (p_i q_j + p_j q_i - p_i p_j - q_i q_j) \geq 0 \tag{2.17}$$

and equality holds if $p_1, \dots, p_n$ and $q_1, \dots, q_n$ are identical distributions.

Fix the points $x_1, \dots, x_n$ in $\mathbb{R}^d$ and a hyperplane H, and call it $\hat{H}$. Suppose the first m points to be on one side of $\hat{H}$, $0 \leq m \leq n$.
We select two points at random according to the p or q distribution, choosing which one with probability 1/2, and connect them if the points are on opposite sides of H. More rigorously: we double our sample space making a copy x'_i for each x_i and we assign the probability $p_i/2$ to the original event x_i, and the probability $q_i/2$ to the copy event x'_i. The connected points are called *homogeneous* if both are original or copies, and *mixed* otherwise.
We define the random variable X_{ij} as:

$$X_{ij} := \mathbb{1}_{\{x_i, x'_j\}} + \mathbb{1}_{\{x'_i, x_j\}} - \mathbb{1}_{\{x_i, x_j\}} - \mathbb{1}_{\{x'_i, x'_j\}}.$$

where $\mathbb{1}_{\{x_i, x'_j\}}$ means that the points x_i, x_j have been chosen and are connected. If $p = \sum_{i=1}^{m} p_i/2$ and $q = \sum_{i=1}^{m} q_i/2$, then the expected number of line segments between mixed pairs minus the expected numbers of line segments between homogeneous pairs, conditioned on the choice of H is:

$$\mathbb{E}\Big[\sum_{i,j=1}^{n} X_{ij} \Big| H = \hat{H}\Big] = p(1-q) + q(1-p) - p(1-p) - q(1-q) \tag{2.18}$$

$$= (p-q)^2 \geq 0 \tag{2.19}$$

and equality holds if and only if $p = q$.
Suppose H is chosen randomly, and consider all the possible r partitions $P_m = (S_{m_1}, S_{m_2})$ of the points $x_1, \dots, x_n$ determined by H, and it this case call H as H_m. Then each partition defines probability distributions $(p^{(m)}, 1-p^{(m)})$ and $(q^{(m)}, 1-q^{(m)})$, such that $p^{(m)} = \sum_{j \in S_{m_1}} p_j$ and $q^{(m)} = \sum_{j \in S_{m_1}} q_j$.
For each of the r possible partitions $P_1, \dots, P_r$ let α_m denote the probability that H determines partition P_m.
Then, using 2.18, the expected number of line segments between mixed pairs minus the expected number of line segments between homogeneous pairs is:

$$\mathbb{E}\Big[\sum_{i,j=1}^{n} X_{ij}\Big] = \sum_{m=1}^{r} \mathbb{E}\Big[\sum_{i,j=1}^{n} X_{ij}\Big| H = H_m\Big] = \sum_{m=1}^{r} \alpha_m (p^{(m)} - q^{(m)})^2 \geq 0. \quad (2.20)$$

With the randomization of H of Lemma 1 (see below), we know that the probability that H intersects the segment (x_i, x_j) is proportional to $\|x_i - x_j\|$. Using this fact, we can count $\mathbb{E}[\sum_{i,j} X_{ij}]$ also in an other way that will allow us to finish the first part of the proof:

$$\mathbb{E}\Big[\sum_{i,j} X_{ij}\Big] = \sum_{i,j=1}^{n} \mathbb{P}(H \cap (x_i, x_j) \neq \emptyset) \cdot \mathbb{E}[X_{ij} | H \cap (x_i, x_j) \neq \emptyset] \quad (2.21)$$

$$\propto \sum_{i,j=1}^{n} \|x_i - x_j\| (p_i q_j + p_j q_i - p_i p_j - q_i q_j). \quad (2.22)$$

From inequality 2.20 we know that this sum is non-negative. Being it proportional to our thesis, we can conclude that also 2.17 is non-negative.

To conclude, we observe that sum 2.20 is non-negative and equals to zero if and only if $(p^{(m)} - q^{(m)}) = 0$ for all m. For each j, there is an m such that $p^{(m)} = p_j$ and $q^{(m)} = q_j$, so equality holds if and only if $p_i = q_i$ for $i = 1, \dots, n$. □

Lemma 1. *Fix $x_1, \dots, x_n \in \mathbb{R}^d$. Select a ball $B \subset \mathbb{R}^d$ with center O that contains the points $x_1, \dots, x_n$, then select a point P uniformly distributed on the surface of B, and a point Q uniformly distributed on the diameter containing OP.*
Choose the hyperplane H such that $Q \in H$ and $H \perp OP$.
Then for each pair (x_i, x_j), the probability that H intersects the segment between x_i and x_j is proportional to $\|x_i - x_j\|$.

Proof. Without loss of generality we assume that the ball B has radius 1 and is centered on the origin O. Moreover we will just study the probability that the hyperplane intersects the segment (x_1, x_2), the other cases are analogous.

The selection of P, Q can be expressed with the parameters $\hat{e}$ and r, where $\hat{e}$ is the direction chosen through unitary vector on the sphere and r is a scalar $r \in (-1, 1)$ that quantifies the distance from the origin and has the sign coherent with the direction of $\hat{e}$. Then the associated hyperplane has the equation

$$x \cdot \hat{e} - r = 0. \tag{2.23}$$

The points x_1, x_2 are separated by the H if and only if it holds that

$$x_1 \cdot \hat{e} - r < 0 < x_2 \cdot \hat{e} - r \quad \text{or} \quad x_2 \cdot \hat{e} - r < 0 < x_1 \cdot \hat{e} - r. \tag{2.24}$$

In particular for fixed $\hat{e}, x_1, x_2$, it will be possible just one of the two cases, but as we will care only about the length of the interval in which r will live, we can assume without loss of generality that $r \in (x_1 \cdot \hat{e}, x_2 \cdot \hat{e})$, that means

$$\mathbb{P}(r \in (x_1 \cdot \hat{e}, x_2 \cdot \hat{e})) = |\hat{e} \cdot (x_2 - x_1)|/2. \tag{2.25}$$

Hence, fixing $\hat{e}$, the probability for r to be such that H intersects the segment is $|\hat{e} \cdot (x_2 - x_1)|/2$.
In conclusion we notice that, writing $x_2 - x_1 = \|x_2 - x_1\|\hat{v}$, with $\|\hat{v}\| = 1$,

$$\mathbb{P}(H \cap (x_1 x_2) \neq \emptyset) = C \int_{\partial B(0,1)} |\hat{e} \cdot \|x_2 - x_1\|\hat{v}| d\hat{e} \tag{2.26}$$

$$\propto \|x_2 - x_1\| \tag{2.27}$$

where C is a constant. This concludes the proof. □

The result holds more generally for powers of the norm for the exponent $\alpha \in (0, 2)$. The proof uses the characteristic functions of random variables and is more technical than the one just shown.

Proposition 2. *If the d-dimensional random variables X and Y are independent with $\mathbb{E}[\|X\|^\alpha] < \infty + \mathbb{E}[\|Y\|^\alpha] < \infty$ for some $0 \leq \alpha < 1$ and $\hat{f}, \hat{g}$ denote their respective characteristic functions, then calling*

$$\mathcal{E}^{(\alpha)}(X, Y) := 2\mathbb{E}[\|X - Y\|^\alpha] - \mathbb{E}[\|X - X'\|^\alpha] - \mathbb{E}[\|Y - Y'\|^\alpha]$$

we have:

(i) For $0 < \alpha < 2$,

$$\mathcal{E}^{(\alpha)}(X, Y) = \frac{1}{C(d, \alpha)} \int_{\mathbb{R}^d} \frac{|\hat{f}(t) - \hat{g}(t)|^2}{|t|_d^{d+1}} dt \tag{2.28}$$

where

$$C(d,\alpha) = 2\pi^{d/2}\frac{\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)}. \tag{2.29}$$

(ii) for $\alpha = 2$,

$$\mathcal{E}^{(2)}(X,Y) = 2|\mathbb{E}[X] - \mathbb{E}[Y]|^2 \tag{2.30}$$

and $\Gamma(\cdot)$ *is the complete gamma function. Thus if* $\alpha \in (0,2)$ *we have* $\mathcal{E}^{(\alpha)}(X,Y) \geq 0$ *with equality to zero if and only if* X *and* Y *are identically distributed, while if* $\alpha = 2$ *we have equality to zero whenever* $\mathbb{E}[X] = \mathbb{E}[Y]$.

Proof. For statement (ii) it is enough to use the independence of the random variables.
For (i), denoting $\overline{f(\cdot)}$ the complex conjugate of $f(\cdot)$, we have:

$$\begin{aligned}|\hat{f}(\cdot) - \hat{g}(\cdot)|^2 &= (\hat{f}(\cdot) - \hat{g}(\cdot))(\overline{\hat{f}(\cdot) - \hat{g}(\cdot)}) \\ &= (1 - \hat{f}(\cdot)\overline{\hat{g}(\cdot)}) + (1 - \overline{\hat{f}(\cdot)}\hat{g}(\cdot)) - (1 - \hat{f}(\cdot)\overline{\hat{f}(\cdot)}) - (1 - \hat{g}(\cdot)\overline{\hat{g}(\cdot)}) \\ &= \mathbb{E}[(2 - (\exp(i(t, X-Y)) + \exp(-i(t, X-Y)))) \\ &\qquad - (1 - \exp(i(t, X - X'))) - (1 - \exp(i, (t, Y - Y')))]\end{aligned}$$

Notice that

$$\exp(i(t, X-Y)) + \exp(-i(t, X-Y)) = 2\cos(t, X-Y)$$

and that

$$\mathbb{E}[\exp(i(t, X-X'))] = \mathbb{E}[\cos(t, X-X') + i\sin(t, X-X')] = \mathbb{E}[\cos(t, X-X')].$$

Combining the two things we arrive to:

$$|\hat{f}(\cdot) - \hat{g}(\cdot)|^2 = 2\mathbb{E}[(1-\cos(t, X-Y)) - (1-\cos(t, X-X')) - (1-\cos(t, Y-Y'))].$$

Integrating over $\mathbb{R}^d$ brings us to:

$$\begin{aligned}&\int_{\mathbb{R}^d} \frac{|\hat{f}(\cdot) - \hat{g}(\cdot)|^2}{\|t\|^{d+\alpha}} dt \\ &= \mathbb{E}\Big[\int_{\mathbb{R}^d} \frac{2(1-\cos(t, X-Y)) - (1-\cos(t, X-X')) - (1-\cos(t, Y-Y'))}{\|t\|^{d+\alpha}} dt\Big].\end{aligned}$$

Proving that

$$\int_{\mathbb{R}^d} \frac{(1-\cos(t,x)}{\|t\|^{d+\alpha}} dt = C(d,\alpha)\|x\|^{\alpha} \tag{2.31}$$

will let us conclude. The result is shown in the lemma below. □

Lemma 2. *For all $x \in \mathbb{R}^d$, if $0 < \alpha < 2$, then:*

$$\int_{\mathbb{R}^d} \frac{1 - \cos(t, x)}{\|t\|^{d+\alpha}} dt = C(d, \alpha)\|x\|^\alpha \tag{2.32}$$

where the integral is meant in the principal value sense: as the $\lim_{\epsilon \to 0} \int_{\mathbb{R}^d \setminus \{\epsilon B + \epsilon^{-1} B^c\}}$ *where B is the unit ball centered on the origin of* $\mathbb{R}^d$.

Proof. We will separate x from the rest in order to have an integral independent of it.
To do so, we apply an orthogonal change of variables that sends $x/\|x\|$ to e_1. The vector t will be sent to $t \mapsto z = (z_1, \dots, z_d)$, so (t, x) will become$z_1 \cdot \|x\|$. Then we apply also the change of variables $s = \|x\| \cdot z$ to get:

$$\int_{\mathbb{R}^d} \frac{1 - \cos(z_1\|x\|)}{\|z\|^{d+\alpha}} dz = \|x\|^\alpha \int_{\mathbb{R}^d} \frac{1 - \cos(s_1)}{\|s\|^{d+\alpha}} dt \tag{2.33}$$

The explicit computations to find the value of the integral, that is $C(d, \alpha)$, can be found in [13]. Anyways, the argument of the integral is positive hence we can already conclude that $\mathcal{E}^{(\alpha)} \geq 0$. □

With this new representation of distance, it is clear that the triangular inequality and symmetry properties also hold true.

Remark 2. *We can see that $\mathcal{E}^{(\alpha)}(X, Y)$ is a weighted L^2 distance between characteristic functions, with weight function $w(t) = \|t\|^{-(d+\alpha)}$. Supposing that the following technical conditions on w hold: w is continuous, $w(t) > 0$ and $\int |\hat{f}(t) - \hat{g}(t)|^2 w(t) dt < \infty$, then it is true that if the weighted L_2 distance between $\hat{f}$ and $\hat{g}$ is rotation invariant and scale equivariant, then $w(t) = const/|t|^{d+\alpha}$. In other words, rotation invariance and scale equivariant imply that the weighted L^2 distance between characteristic functions is the energy distance.*

Equivariance means that:

$$\int |\hat{f}(at) - \hat{g}(at)|^2 w(t) dt = |a| \cdot \int |\hat{f}(t) - \hat{g}(t)|^2 \quad \text{for } a \geq 0.$$

2.1.2 Implementation

For our experiments we use as negative definite kernel the function $h(\cdot)$ defined as follows:

$$h(x) = \sqrt{a^2 + \|x\|^2} - a \tag{2.34}$$

for some given constant $a \geq 0$.
The following theorem from Schoenberg [10], Micchelli [5], Turinici [16] assures us that the choice of the function is legit.

Theorem 3. *For any $a \geq 0, \alpha \in (0,1)$, the kernel $h(x) = (a+|x|^2)^\alpha$ is negative definite and can be expressed explicitly as a Gaussian mixture. In particular is true for $\sqrt{a+x^2}$.*

Then given two probability measures η_1, η_2, we define the kernel distance:

$$d(\eta_1,\eta_2)^2 = \iint h(|x-y|)(\eta_1-\eta_2)(dx)(\eta_1-\eta_2)(dy). \tag{2.35}$$

In our case both η_1 and η_2 are sums of Dirac masses:

$$\eta_1 = \sum_{k=1}^{K} \alpha_k \delta_{x_k}, \quad \eta_2 = \sum_{j=1}^{J} \beta_j \delta_{y_j} \tag{2.36}$$

with $K, J < \infty$, $\alpha_i, \beta_j > 0$. So the formula of the distance becomes:

$$\begin{aligned} d(\eta_1,\eta_2)^2 = &\sum_{k=1}^{K}\sum_{j=1}^{J} \alpha_k\beta_j h(|x_k-y_j|) \\ &- \frac{1}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K} \alpha_k\alpha_{k'} h(|x_k-x_{k'}|) - \frac{1}{2}\sum_{j=1}^{J}\sum_{j'=1}^{J} \beta_j\beta_{j'} h(|y_j-y_{j'}|) \end{aligned} \tag{2.37}$$

2.1.3 Algorithm

As will be better discussed in Chapter 3, we will be interested in the following problem:
Given a sum of Dirac distribution $\mu = \frac{1}{N_K}\sum_{k=1}^{N_K} \delta_{x_k}$, find the distribution sum of K Dirac $\nu = \frac{1}{K}\sum_{k=1}^{K} \delta_{y_k}$ that minimizes the Kernel Distance.
From the work of Turinici [17], we know that this minimization problem admits at least one solution.
To find the points $y_1, \ldots, y_k$ in which the distribution ν is centered, we use the following algorithm:

```
Algorithm:
Begin Procedure:
    • set $a = 10^{-6}$
    • define $h$ of 2.37 $h(x) = \sqrt{a^2 + |x|^2} - a$.
    • while (max iteration not reached) do:
        – sample $Z = z_1, \ldots, z_K$ randomly
        – compute the global loss:
              $$L(Z) = d\left(\frac{1}{N_K}\sum_{k=1}^{N_K}\delta_{x_k}, \frac{1}{K}\sum_{j=1}^{K}\delta_{z_j}\right)^2 \qquad (2.38)$$
        – backpropagate the loss $L(Z)$ in order to minimize $L(Z)$
        – update $Z$.
    end while
end procedure.
```

For the backpropagation we used the Adam optimizer, a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments.

From the work of Turinici [15] we know that the Loss function defined with the kernel distance is convex and hence the problem has good convergence properties.

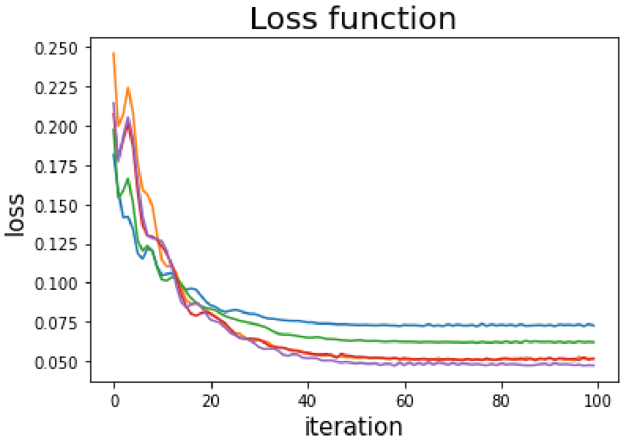

Figure 2.1: Graphs of loss function minimization for keyword search for 5 phrases.

2.2 Geodesic distance on the hypersphere

We developed another strategy by exploiting the fact that all word vectors can be normalized. In this situation, the vectors lie on the surface of the unit hypersphere and we can use a more appropriate kernel, as done for example in [1].
Specifically, we replace the h kernel of 2.34 with the geodesic distance, which is the length of the maximal arc passing between two points:

$$g(x, y) = \arccos(x \cdot y) \tag{2.39}$$

and the distance between two distributions of the kind of sum of Diracs $\sum \delta_{x_i}$ has the expression:

$$\begin{aligned} d(\eta_1, \eta_2)^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_k \beta_j g(x_k, y_j) - \frac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \alpha_k \alpha_{k'} g(x_k, x_{k'}) \\ - \frac{1}{2} \sum_{j=1}^{J} \sum_{j'=1}^{J} \beta_j \beta_{j'} g(y_j, y_{j'}) \end{aligned} \tag{2.40}$$

We refer to future work [11] for additional theoretical results on the hypothesis required to ensure that the kernel in 2.39 leads to a proper metric.

2.3 k-means clustering

In this section we explain what the k-means clustering algorithm is [4], as we try to use it for an experiment in Chapter 3.

Given a set of observations $(x_1, \ldots, x_n)$, where each observation is a $d-$dimensional real vector, the $k-$means clustering algorithm tries to partition the observations into k $(\leq n)$ sets $S = \{S_1, \ldots, S_K\}$ so as to minimize the within-cluster sum of squares. Formally, the goal of the algorithm is to find:

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg\min_S \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{2.41}$$

where μ_i is the mean of points in S_i.

We will use this method with both Euclidean and geodesic kernels. In the former case we will use the package `scikit-learn`, and for the latter the package implemented by [7].

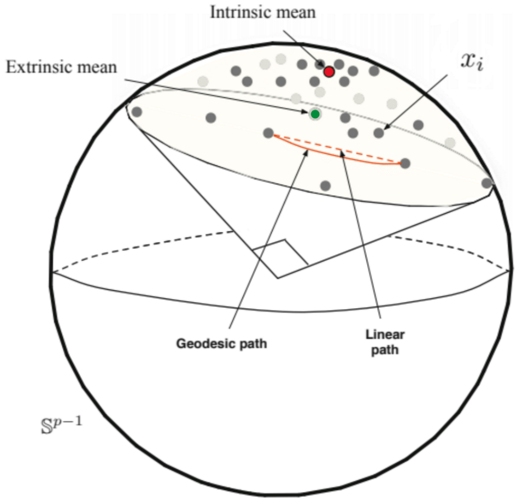

Figure 2.2: Representation from [1] of the unit vectors lying on the sphere. In orange there are the geodesic and linear paths connecting two vectors. In green there is the standard mean, done with the euclidean norm. In red there is the mean obtained using the geodesic distance.

Chapter 3

Experiments

The purpose of the experiments we present in this chapter is to extract from each sentence in our datasets three keywords that express the meaning of the sentence. To handle this task, we focus primarily on the Kernel Distances studied earlier.
We used two datasets:

- the first one is a collection of tweets of the most diverse topics, from politics to movies (taken from *https://data.world/a2liz/favorited-tweets*);
- the second one is the original version of "Alice in Wonderland".

The latter dataset is used to have sentences in a larger context, as we sought better performance by also employing the environment in which our sentences lived.

One of the main difficulties of this task is that so far there is no metric that can calculate the goodness of the method, since there are several valid choices for keywords, and different people may choose different words. Therefore, all observations are based on the opinions of the authors who observed the results one by one.

The experiments were carried on with the GloVe embedding that transformed each word in a vector.
We used the 50-dimensional embedding vectors of the words provided by the web page of the creators (`https://nlp.stanford.edu/projects/glove/, 6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download`).

The code can be found in the Github page:
`https://github.com/ceciliasecchi/MasterThesis`

3.1 Pre-Process of the dataset

Before using our data sets, we pre-processed them to preserve only their essence in order to facilitate our task.
We did the following operations:

- text was divided in sentences;
- stopwords from `gensim` [9] were removed, but also adverbs and conjunctions such as `'and','but','this','the','how','for','like','while','instead'` as they don't carry meaningful semantic information;
- words of less than 3 characters were removed;
- punctuation was removed;
- all the characters were made lower;
- short sentences (of less than 5 words) were removed from the dataset.

In conclusion we obtained this kind of transformation:

"I wonder if I shall fall right *through* the earth!" → "wonder shall fall right through earth"

"It was all very well to say "Drink me," but the wise little Alice was not going to do *that* in a hurry." → "drink wise little alice going hurry".

3.1.1 Normalization

Since we observed that almost all the words in the dataset have a relatively similar norm between 4 and 6 (see the left graph of 3.1), for a batch of experiments we normalized them and exploited their distribution over the unit sphere. In particular, for a series of experiments, we used the geodesic distance of the sphere as the kernel between the vectors.

We studied the distribution of words on the sphere. Since they had size 50, we could not visualize them, so we studied the maximal geodesic distance between each word. The result is shown in the histogram to the right of 3.1. We can see that most of the sentences have all their component words within a relatively small section of the sphere, since the maximum degree is roughly $pi/2$.

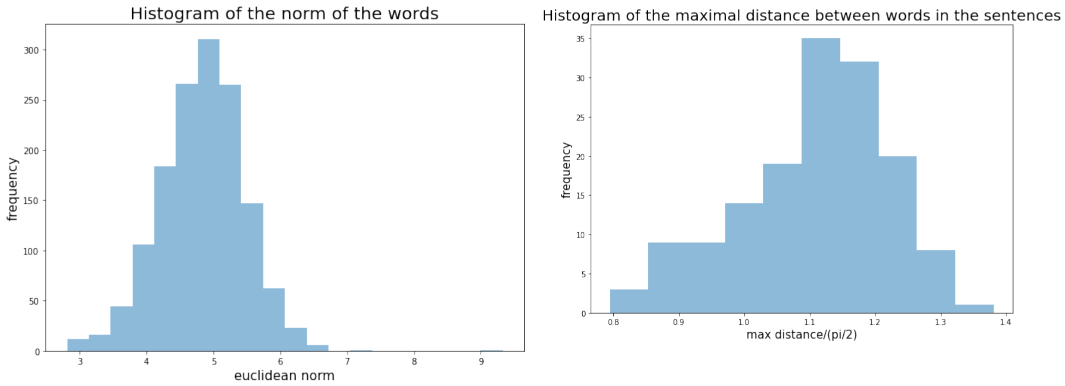

Figure 3.1: On the left: histogram of the euclidean norm of the words of the Tweeter dataset
On the right: histogram of the maximal arch-distance of the words in each sentence.

3.2 First Experiment: the Average

As a first experiment, we worked with the Tweets dataset and we tried to find the essence of the sentences doing the average of the vectors that compose it, finding the Average Vector (AV). From the AV we searched for the 3 nearest word-vectors in the set of the words of the dataset.
We found one neighboor word at a time, looking for the 3 word-vector that maximize the cosine similarity to the AV in the dataset.
We tried some variations to this method weighting in different ways the words that composed the sentences to obtain the AV. In particular we used:

- Tf-Id (term frequency–inverse document frequency) for the dataset of Alice in Wonderland
- Inverse of frequency coefficients obtained by a large dataset found in internet, independent of the dataset from which the sentences under consideration came from.

The idea of the latter weight is to give more importance to rare words as they will be peculiar to the sentence and bring semantic meaning.

3.2.1 Results

In this subsection there are the sentences and the associated triplets found with the three methods.

Social Security and Medicare are not ”Entitlements” that can be renegotiated. The term is ”DEFERRED PAY.” I've paid into the system my entire life, and so have you. Anyone stealing your deferred pay needs to be voted out.
[PAY, WIFE, CHARITABLE]
[PAY, WIFE, PAY]
[RENEGOTIATED, TERM, CHRIS]

’You let a weird guy whisper in your ear and you didn't trust your daughter.’ –my 7-year-old girl.
[LOVE, POOR, SCARIEST]
[LOVE, KNOWING, POOR]
[WHISPER, PRETEND, WHISPER]

It's interesting watching this with them because I thought Gollum might scare them but they mostly just get upset when he's mistreated.
[AFRAID, SCARIEST, APOCALYPTIC]
[AFRAID, SCARIEST, ENTIRE]
[GOLLUM, UPSET, NICE]

I first saw it as Danny DeVito looking outside the window, disappointed to see a topless Woman wearing the same speedo he did.
[LOOKING, DAY, GIVE]
[LOOKING, DAY, GIVE]
[SPEEDO, DEVITO, CEREAL]

Marvel should make a mockumentary about Loki as Odin and the behind the scenes drama that went on while Loki was trying to produce and direct his play. i wanna see tantrums about sets, actors and interviews with asgardians who are ya we know it's Loki
[CHARACTER, MOVIE, IMPORTANTLY]
[LOKI, CRAZIEST, LAYING]
[ASGARDIANS, LOKI, ARAGORN]

A couple years ago, a very nice young girl was being bullied for wearing a feminist t-shirt in a photo. I offered her words of support. Today, in Sao Paulo, she proudly handed me a graphic novel she wrote. I love comics.
[ONE, WELL, WHY]

[ONE, WELL, WHY]
[PAULO, SAO, MEDIA]

I'm sure sexual assault charges in the military would decrease if men didn't serve

[TAKING, RETROSPECT, MATT]
[TAKING, RETROSPECT, SEXUAL]
[ICREASED, TAKING, CHARGES]

We need to fix this country's infrastructure. It should be a major priority for the president and Congress.
[NEEDS, INFRASTRUCTURE, CHILDREN]
[PRIORITY, NEEDS, CHILDREN]
[COUNTRY, STABLE, ABSOLUTELY]

My wife makes more money than me. Claims it means she gets to set the thermostat in the house.
[TURN, CRITIQUE, FAMILY]
[TURN, CRITIQUE, FAMILY]
[THERMOSTAT, FEVER, KNOWN]

Dear Brave Airport Lady, if you ever see this, just know that although I'm sorry you went through this (esp during the holidays), I hope that, one day, you look back on this as a defining moment in your life. Never settle. Follow your heart. You are a goddess.
[COME, QUESTION, INTELLIGENT]
[COME, QUESTION, INTELLIGENT]
[ESP, SOUNDS, PSA]

3.2.2 Observations

We were not particularly satisfied of the results, and this was likely due to the fact that the average of a sentence loses too much information.
Moreover the research of the keywords as the nearest vectors to the AV makes them synonyms. As a sentence can have more semantic areas, this doesn't allow to preserve the whole spectrum of arguments.
We can observe that the plain average and the tf-idf average produce almost identical results, while the independent frequencies give very different results. We cannot say that a method is better than the others, as the performance change from case to case. For example for the first sentence

"We need to fix this country's infrastructure. It should be a major priority for the president and Congress"
The plain AV is better than the independent frequencies as they have as keywords "needs" and "infrastructure", while the second doesn't center the meaning of the sentence returning "country, stable, absolutely".
While fot the sentence:
"I first saw it as Danny DeVito looking outside the window, disappointed to see a topless Woman wearing the same speedo he did" The third method wins as it returns "speedo" and "devito" while the first one is too general with its "looking,day,give".

3.3 Second experiment: k-mean clustering

Our second approach, related to the first experiment, was to generalize the mean into a k-mean clustering for $k = 3$. We thought that this would avoid the risk of choosing 3 synonyms as keywords, since each centroid would synthesize different semantic areas of the sentence.

We also tried a geometric k-mean clustering. For this experiment, we normalized the word vectors so that they were all embedded in the unit sphere. With this setup, we used the geodesic distance as the distance for the algorithm and no longer the Euclidean distance.

In the standard k-mean clustering for each centroid the corresponding keyword was found using the maximizer of the projection on the centroid, while in the second case the geodesic minimizer.

3.3.1 Results

In this section we show the results obtained by the two methods. The first triplet is the original k-mean clustering, while the second is the geodesic clustering.

Social Security and Medicare are not "Entitlements" that can be renegotiated. The term is "DEFERRED PAY." I've paid into the system my entire life, and so have you. Anyone stealing your deferred pay needs to be voted out.
[PAY, MAKING, DEFERRED]
[MAKING, PAY, SECURITY]

'You let a weird guy whisper in your ear and you didn't trust your daughter.' –my 7-year-old girl.
[MOTHER, BIT, TRUST]
[WHISPER, WEIRD, MOTHER]

It's interesting watching this with them because I thought Gollum might scare them but they mostly just get upset when he's mistreated.
[REALLY, MISTREATED, GOLLUM]
[WATCHING, INTERESTING, SCARE]

I first saw it as Danny DeVito looking outside the window, disappointed to see a topless Woman wearing the same speedo he did.
[LOOKING, DANNY, SPEEDO]
[DANNY, SAW, WEARING]

I have but one request this Christmas, guys. Take the #WarOnChristmas hashtag and fill it with accounts of the horrors of battling elves across snowy fields and firing anti-aircraft missiles at sleds streaking overhead.
[COME, SLEDS, MISSILES]
[FIRING, LIFE, STREAKING]

Marvel should make a mockumentary about Loki as Odin and the behind the scenes drama that went on while Loki was trying to produce and direct his play. I wanna see tantrums about sets, actors and interviews with asgardians who are ya we know it's Loki
[LOKI, MAKING, DRAMA]
[LOKI, ASGARDIANS, MAKING]

A major cause for celebration and dancing in the streets: Henry the Hatter to open in Eastern Market on Friday
[CELEBRATION, TODAY, HENRY]
[OPEN, DANCING, MAJOR]

A couple years ago, a very nice young girl was being bullied for wearing a feminist t-shirt in a photo. I offered her words of support. Today, in Sao Paulo, she proudly handed me a graphic novel she wrote. I love comics.
[PAULO, ONE, BOOK]
[TODAY, WOMAN, BOOK]

I'm sure sexual assault charges in the military would decrease if men didn't

serve
[ASSAULT, TAKE, SEXUAL]
[KNOW, SERVE, ASSAULT]

We need to fix this country's infrastructure. It should be a major priority for the president and Congress.
[NEEDS, PRESIDENT, COUNTRYS]
[NEEDS, COUNTRYS, CONGRESS]

My wife makes more money than me. Claims it means she gets to set the thermostat in the house.
[MAKING, THERMOSTAT, WIFE]
[CLAIMS, MAKING, THERMOSTAT]

Dear Brave Airport Lady, if you ever see this, just know that although I'm sorry you went through this (esp during the holidays), I hope that, one day, you look back on this as a defining moment in your life. Never settle. Follow your heart. You are a goddess.
[COME, DEAR, ESP]
[DEAR, COME, FOLLOW]

3.3.2 Observations

We are more satisfied with these results because in most cases, with both methods, the words are related to the subject of the sentence. For example, in the first sentence ”Social Security and Medicare...” the triplet ”pay, making, deferred” is better than that found with the first experiment ”pay, wife, charitable.”

It cannot be said which method is better between the two k-mean clustering: in some sentences one is better than the other, but the roles often interchange. For example, in the sentence ”You let a strange guy...” the geodesic k-mean is better because ”whisper, strange, mother” are all important words related to the sentence, while ”mother, bit, trust” contains ”bit” which is useless.
While in ”It is interesting to watch this with them...” the first triplet ”really, mistreated, gollum” contains ”gollum” and ”mistreated” which are better and more precise about the meaning of the sentence, rather than ”watching, interesting” found with the geodesic k-mean.

3.4 Third experiment: Kernel Distance

As third experiment we approached the problem in a totally different way using the theoretical results of chapter 2.
We imagined that each Thought/Concept has a certain distribution to be expressed, and that each word that composes a sentence, is a sample from that distribution of the concept.
In this way we can see that the word-vectors that compose a sentence as a sum of Dirac distributions that approximate the Thought. The goal is then to search a sum of 3 Dirac that compresses and approximate this distribution.

In formulas, if the sentence S is composed by the N words $S_1, \ldots, S_N$, we define

$$\mu_S = \sum_{i=1}^{N} \frac{1}{N} \delta_{S_i}.$$

The goal is to find the distribution

$$\mu_C = \sum_{i=1}^{3} \frac{1}{3} \delta_{C_i}$$

that compresses the information of μ_S minimizing the Energy Distance:

$$\mu_C = \underset{\mu_3 = \sum_{i=1}^{3} \delta_{X_i}}{\arg\min} \; d(\mu_S, \mu_3).$$

As the optimization algorithm (described in chapter 2) returns 3 compression-vectors X_1, X_2, X_3 that don't necessarily correspond to word-vectors of our vocabulary, we have to search for each compression-vector the word-vector nearest to it. We chose as criterion to find these words the following:
for each X_i we found the word C_i that maximizes the projection:

$$C_i = \underset{C \in \text{Vocabulary}}{\arg\max} \frac{X_i \cdot C}{\|X_i\| \|C\|}$$

Remark 3. *The sentences in the dataset vary widely in length, from 6 to 28 significant words. To homogenize the length and computation time for this optimization experiment, we made them all 30 words by adding or selecting words, chosen with uniform probability from the set of words that make up the sentence.*

3.4.1 Results

We report the triplets obtained by the method described above, using the h function as the kernel for the statistical distance: $h(x, y) = \sqrt{c^2 + |x - y|^2} - c$.

Social Security and Medicare are not ”Entitlements” that can be renegotiated. The term is ”DEFERRED PAY.” I’ve paid into the system my entire life, and so have you. Anyone stealing your deferred pay needs to be voted out.
[DEFERRED, ANYONE, SECURITY]

’You let a weird guy whisper in your ear and you didn’t trust your daughter.’ –my 7-year-old girl.
[WHISPER, WEIRD, MOTHER]

It’s interesting watching this with them because I thought Gollum might scare them but they mostly just get upset when he’s mistreated.
[INTERESTING, MISTREATED, SCARE]

I first saw it as Danny DeVito looking outside the window, disappointed to see a topless Woman wearing the same speedo he did.
[DEVITO, WEARING, COMING]

I have but one request this Christmas, guys. Take the #WarOnChristmas hashtag and fill it with accounts of the horrors of battling elves across snowy fields and firing anti-aircraft missiles at sleds streaking overhead.
[WEED, OUT, WHOS]

Marvel should make a mockumentary about Loki as Odin and the behind the scenes drama that went on while Loki was trying to produce and direct his play. I wanna see tantrums about sets, actors and interviews with asgardians who are ya we know it’s Loki.
[LOKI, ACTORS, SHOW]

A major cause for celebration and dancing in the streets: Henry the Hatter to open in Eastern Market on Friday
[CELEBRATION, CAUSE, OPEN]

A couple years ago, a very nice young girl was being bullied for wearing a feminist t-shirt in a photo. I offered her words of support. Today, in Sao Paulo, she proudly handed me a graphic novel she wrote. I love comics.
[HIM, GRAPHIC, TODAY]

I'm sure sexual assault charges in the military would decrease if men didn't serve
[ASSAULT, NEED, CHARGES]

We need to fix this country's infrastructure. It should be a major priority for the president and Congress.
[CONGRESS, FIX, INFRASTRUCTURE]

My wife makes more money than me. Claims it means she gets to set the thermostat in the house.
[WIFE, MEANS, MAKES]

3.4.2 Observations

We weren't particularly satisfied of the results, so we investigated them through some statistics. For this reason we computed the goodness of our approaches studying the following metrics:

- $d(S, C_1)$: the distance between the sentence to the 3 original compression vectors
- $d(S, C_P)$: the distance between the sentence and the projected compression vectors
- $d(C_P, C_1)$: the distance between the original vectors and the projected ones
- $d(S, C_R)$: the distance between the sentence and 3 random words taken from the sentence.

Results showed that the projection onto the compression-word destroyed the proximity obtained through the Energy distance, as we can see from the data of table 3.1: from an approximate distance of 0.3 we pass to a distance of 0.6, that is equal to the distance of the sentence to a random triplet obtained from the sentence (0.6).

3.5 Fourth experiment

As a last experiment we tried to output the triplet of words T_1, T_2, T_3 within the sentence whose distribution μ_T minimizes the kernel distance. We tried

$d(S, C_1)$	$d(S, C_P)$	$d(S, C_R)$	$d(C_P, C_1)$
0.267	0.620	0.669	0.671
0.284	0.622	0.600	0.660
0.313	0.587	0.546	0.669
0.265	0.588	0.589	0.668
0.289	0.570	0.636	0.656

Table 3.1: Statistics obtained from some tweets of the dataset.

with two different kernels:

$$h(x, y) = \sqrt{a^2 + |x - y|^2} - a \quad \text{and} \quad g(x, y) = \arccos{(x \cdot y)}.$$

The advantage of this method is that certainly the triplet is strongly related to the sentence, but it has the downside of not being able to find words outside the sentence that might better describe its meaning.

The idea of looking for the best triplet outside the dataset is not feasible because of the computational cost, so we tried to find the best triplet by searching in a larger set of words that are somehow related to our sentence. Since it was not easy to find this kind of set with the Tweets dataset, we switched to the "Alice in Wonderland" dataset and used all the words m within the sentence page as the context set $\{CT\}$.

In particular, using the same notation of the previous section, after having found the compression distribution μ_X with the optimization algorithm, we approximate μ_X with the distribution $\mu_M = \frac{1}{3}\sum_{i=1}^{N} \delta_{M_i}$ centered on the word vectors M_i found within the set of contexts $\{CT\}$ using the following algorithm:

```
Algorithm:
$\mu_S = \frac{1}{N} \sum_{i=1}^{N} \delta_{S_i}$
$X_1, X_2, X_3 = \arg\min_{X_1,X_2,X_3} d(\sum_{i=1}^{3} \frac{1}{3}\delta_{X_i}, \mu_S)$
$\mu_X = \frac{1}{3}\delta_{X_1} + \frac{1}{3}\delta_{X_2} + \frac{1}{3}\delta_{X_3}$

    • $M_1 = \arg\min_{M \in \{CT\}} d(\delta_m, \mu_C)$
    • $M_2 = \arg\min_{M \in \{CT\}} d(\frac{1}{2}\delta_{M_1} + \frac{1}{2}\delta_m, \mu_C)$
    • $M_3 = \arg\min_{M \in \{CT\}} d(\frac{1}{3}\delta_{M_1} + \frac{1}{3}\delta_{M_2} + \frac{1}{3}\delta_M, \mu_C)$
return $M_1, M_2, M_3$
end procedure.
```

3.5.1 Glove Results

The triplet with the kernel h is shown below. Since the results obtained with geodesics are almost identical, they are not reported.

Social Security and Medicare are not ”Entitlements” that can be renegotiated. The term is ”DEFERRED PAY.” I’ve paid into the system my entire life, and so have you. Anyone stealing your deferred pay needs to be voted out.
[ENTITLEMENTS, PAID, ANYONE]

’You let a weird guy whisper in your ear and you didn’t trust your daughter.’ –my 7-year-old girl.
[WEIRD, GUY, GIRL]

it’s interesting watching this with them because I thought Gollum might scare them but they mostly just get upset when he’s mistreated.
[INTERESTING, WATCHING, UPSET]

I have but one request this Christmas, guys. Take the #WarOnChristmas hashtag and fill it with accounts of the horrors of battling elves across snowy fields and firing anti-aircraft missiles at sleds streaking overhead.
[ANTIAIRCRAFT, MISSILES, SLEDS]

Marvel should make a mockumentary about Loki as Odin and the behind the scenes drama that went on while Loki was trying to produce and direct his play. I wanna see tantrums about sets, actors and interviews with asgardians who are ya we know it’s Loki

[TANTRUMS, ACTORS, LOKI]

I first saw it as Danny DeVito looking outside the window, disappointed to see a topless Woman wearing the same speedo he did.
[WINDOWS, TOPLESS, WOMAN]

A major cause for celebration and dancing in the streets: Henry the Hatter to open in Eastern Market on Friday
[CELEBRATION, HENRY, HATTER]

A couple years ago, a very nice young girl was being bullied for wearing a feminist t-shirt in a photo. I offered her words of support. Today, in Sao Paulo, she proudly handed me a graphic novel she wrote. I love comics.
[WEARING, PHOTO, NOVEL]

I'm sure sexual assault charges in the military would decrease if men didn't serve.
[SURE, MEN, SERVE]

We need to fix this country's infrastructure. It should be a major priority for the president and Congress.
[COUNTRYS, PRESIDENT, CONGRESS]

My wife makes more money than me. Claims it means she gets to set the thermostat in the house.
[MAKES, CLAIMS, MEANS]

3.5.2 Alice Results

Here are the results obtained using the Alice dataset, specifically studying the first page of the book.
The first triplet is the best one with the words in the sentence, the second and third were found by the method described above using $h(x,y) = \sqrt{c^2 + |x-y|^2} - c$ and $g(x,y) = \arccos(x \cdot y)$ as the kernel for statistical distance.

Next to the triplets are the kernel distances. $d(S, C_T)$ is the distance between the sentence and the best triplet, $d(S, C_1)$ is the distance between the sentence and the minimizing vector triplet, $d(S, C_P)$ is the distance between the sentence and the word triplet closest to the minimizing vector triplet.

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it,

[SITTING ,BOOK, READING] $d(S, C_T) = 0.5872$

[HAVING, NOTICED, ALICE] $d(S, C_1) = 0.6556$ $d(S, C_P) = 0.6805$

[HAVING, NOTICED, ALICE] $d(S, C_1) = 0.7225$ $d(S, C_P) = 0.7154$

So she was considering in her own mind (as well as she could, for the hot day made her feel -very- sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her

[HEAR, RABBIT, ITSELF] $d(S, C_T) = 0.654$

[MAKING, SOMEBODY, RABBIT] $d(S, C_1) = 0.646$ $d(S, C_P) = 0.706$

[MAKING , FORTUNATELY, RABBIT]$d(S, C_1) = 0.688$ $d(S, C_P) = 0.749$

nor did Alice think it so very much out of the way to hear the Rabbit say to itself,

[HEAR, RABBIT, ITSELF] $d(S, C_T) = 0.654$

[THINK, STUPID, READING] $d(S, C_1) = 0.679$ $d(S, C_P) = 0.702$

[ITSELF, AFTERWARDS, NOTICED] $d(S, C_1) = 0.727$ $d(S, C_P) = 0.847$

But when the Rabbit actually -took a watch out of its waistcoat-pocket-, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge

[RABBIT, WATCH, POP] $d(S, C_T) = 0.578$

[WHEN, NOTICED, ALICE] $d(S, C_1) = 0.614$ $d(S, C_P) = 0.688$

[TIME, FORTUNATELY, ALICE] $d(S, C_1) = 0.675$ $d(S, C_P) = 0.735$

The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well

[MOMENT, THINK, FALLING] $d(S, C_T) = 0.664$

[DOWN, SOMEBODY, RABBIT] $d(S, C_1) = 0.687$ $d(S, C_P) = 0.713$

[SUDDENLY, SOMEBODY, RABBIT] $d(S, C_1) = 0.751$ $d(S, C_P) = 0.764$

Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her and to wonder what was going to happen next
[EITHER, LOOK, WONDER] $d(S, C_T) = 0.679$
[GOING, FORTUNATELY, NOTICED] $d(S, C_1) = 0.687$ $d(S, C_P) = 0.752$
[PLENTY, FORTUNATELY, NOTICED] $d(S, C_1) = 0.785$
$d(S, C_P) = 0.812$

Then she looked at the sides of the well, and noticed that they were filled with cupboards and book-shelves
[FILLED, CUPBOARDS, BOOKSHELVES] $d(S, C_T) = 0.658$
[NOTICED, AFTERWARDS, EITHER] $d(S, C_1) = 0.669$ $d(S, C_P) = 0.726$
[FILLED, AFTERWARDS, EITHER] $d(S, C_1) = 0.688$ $d(S, C_P) = 0.786$

But to her great disappointment it was empty: she did not like to drop the jar for fear of killing somebody underneath, so managed to put it into one of the cupboards as she fell past it
[KILLING, SOMEBODY, MANAGED] $d(S, C_T) = 0.593$
[DOWN, AFTERWARDS, TUNNEL] $d(S, C_1) = 0.640$ $d(S, C_P) = 0.701$
[PAST, AFTERWARDS, EITHER] $d(S, C_1) = 0.685$ $d(S, C_P) = 0.800$

3.5.3 Observations

The results obtained from the best triplet within the sentence are quite good, and the fact that numerically it is the method that minimizes the Kernel Distance compared to the other methods seems to confirm the feeling.

It can be observed that the procedure for finding the trio of words closest to the trio of vectors it minimizes is quite good, since it does not increase the distance from the sentence too much.
In addition, the fact that often the keywords found with the square root kernel and those found with the geodesic kernel are the same can be explained by the observation made earlier: the word-vectors are indeed close in the sphere, which makes the two kernels similar.

Conclusions

In this paper, we applied some statistical distances to NLP to obtain key words of sentences from different datasets.
We studied the GloVe method that allowed us to transform words into vectors, and we studied the mathematical properties of the kernel distances used.
Finally, we tried different techniques to extract keywords from some datasets. Probably the best algorithm is the one that returns the triplet of words within the sentence that minimizes the energy distance. This method provides a simple and versatile procedure that in a known time returns the searched keywords. The time complexity is $\mathcal{O}(n^3)$, where n is the length of the sentence, but this is not a problem since the sentences in the data set used are on average 12-13 significant words long.

However, the procedure needs to be improved because we have not been able to find a metric that allows us to objectively calculate the goodness of our keywords. This is not an easy task because the choice of keywords is subjective and multiple words may be suitable. So far the procedure has been judged by our taste.

In addition, improvements could be made with proper weighting of words in the sentence and by trying other kernels within the chosen statistical distance.

Bibliography

[1] Nicolas Courty, Thomas Burger, and Pierre-François Marteau. Geodesic analysis on the gaussian rkhs hypersphere. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 299–313. Springer, 2012.

[2] Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

[3] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[4] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[5] Charles A Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, pages 143–145. Springer, 1984.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, et al. Geomstats: a python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[9] Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[10] Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, pages 811–841, 1938.

[11] C. Secchi and G. Turinici. Geodesic statistical distances on spheres. in preparation, 2022.

[12] Gábor J Székely and Maria L Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[13] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[14] Gabor J Szekely, Maria L Rizzo, et al. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–184, 2005.

[15] Gabriel Turinici. Radon–sobolev variational auto-encoders. *Neural Networks*, 141:294–305, 2021.

[16] Gabriel Turinici. Unbiased metric measure compression. *doi: 10.5281/zenodo.5705389*, 2021.

[17] Gabriel Turinici. Algorithms that get old: the case of generative algorithms. *arXiv preprint arXiv:2202.03008*, 2022.

[18] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–37, 2008.

[19] AA Zinger, Ashot V Kakosyan, and Lev B Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 59(4):914–920, 1992.