Università degli Studi di Milano

Data Science and Economics
Documents Similarity in Large Datasets

Ivo Bonfanti
XXXXXX

Abstract

This project aims at detecting similarity among pairs of textual documents within a big data environment. As a first thing, a large dataset collecting abstracts from the biomedical domain is tokenized, lemmatized and curated for stop words removal. Then, for each couple of articles, the Jaccard similarity is computed and the number of similar pairs is investigated for different thresholds. Among all, those most similar are furtherly analyzed to asses the accuracy of the result. Finally, to reduce the computational load, an alternative approach, based on Locality Sensitive Hashing, is also provided. While the analysis is performed over a small fraction of the dataset, the proposed solutions are easily scalable. Indeed, they are conducted in a cluster-computing framework as PySpark, specific for handling big data.

ACADEMIC YEAR 2022-2023

Contents

1 Data Set

The employed dataset is called MeDAL[1], standing for medical dataset for abbreviation disambiguation for natural language. It is a collection of 14,393,619 articles from the biomedical domain, curated for abbreviation disambiguation and made publicly available on Kaggle. The dataset is organized as follow,

Table 1: First records of the MeDAL dataset.

Text	**Label**	**Location**
alphabisabolol has a primary antipeptic action depending on dosage which is not caused by an alterat...	56	substrate
a report is given on the recent discovery of outstanding immunological properties in ba ncyanoethyle...	24|49|68|113|137|172	carcinosarcoma, recovery, reference, recovery, after, plaque
the virostatic compound nndiethyloxotetradecylimidazolidinylethylpiperazinecarboxamidehydrochloride ...	55	substrate
rmi rmi and rmi are newly synthetized nrdibenzobfoxepinylnmethylpiperazinemaleates which show intere...	25|82|127|182|222	compounds, compounds, inhibitory, lethal

The first column reports the abstracts, the second the indexes referring to the location within the text of the inserted abbreviations, while the third the

[1]MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. Zhi Wen1, Xing Han Lu1, Siva Reddy

actual substituted words. For the purpose of this analysis only the texts of the articles are useful, therefore the last two columns are dropped. Moreover, the analysis is conducted considering only the first 1000 records. However, the proposed solution is easily scalable to the whole dataset, as explained further below.

2 Preprocessing

For enhancing comparability, the articles must first undergo to some sort of transformation. At first they are tokenized by splitting the strings, therefore converting the texts into set of words. Among them those considered 'stop' words, which are those most commonly used thus not carrying any specific information about the text, are filtered out. Avoiding this step would make the comparison really hard, given that most documents would be classified as similar. Moreover, all the words are converted to lower case and lemmatized. This last technique consists in turning each word in its base form, since the machine would otherwise recognize as different tokens of the kind 'going' and 'go', which would be inconvenient for the similarity detection task. For the sake of clarity an abstract and its transformed version are reported here below,

[*Alphabisabolol has a primary antipeptic action depending on dosage which is not caused by an alteration of the phvalue the proteolytic activity of pepsin is reduced by percent through addition of bisabolol in the ratio of the antipeptic action of bisabolol only occurs in case of direct contact in case of a previous contact with the ATP the inhibiting effect is lost.*]

[*'alphabisabolol', 'primary', 'antipeptic', 'action', 'depend', 'dosage', 'cause', 'alteration', 'phvalue', 'proteolytic', 'activity', 'pepsin', 'reduce', 'percent', 'addition', 'bisabolol', 'ratio', 'antipeptic', 'action', 'bisabolol', 'occurs', 'case', 'direct', 'contact', 'case', 'previous', 'contact', 'atp', 'inhibit', 'effect', 'lose'*]

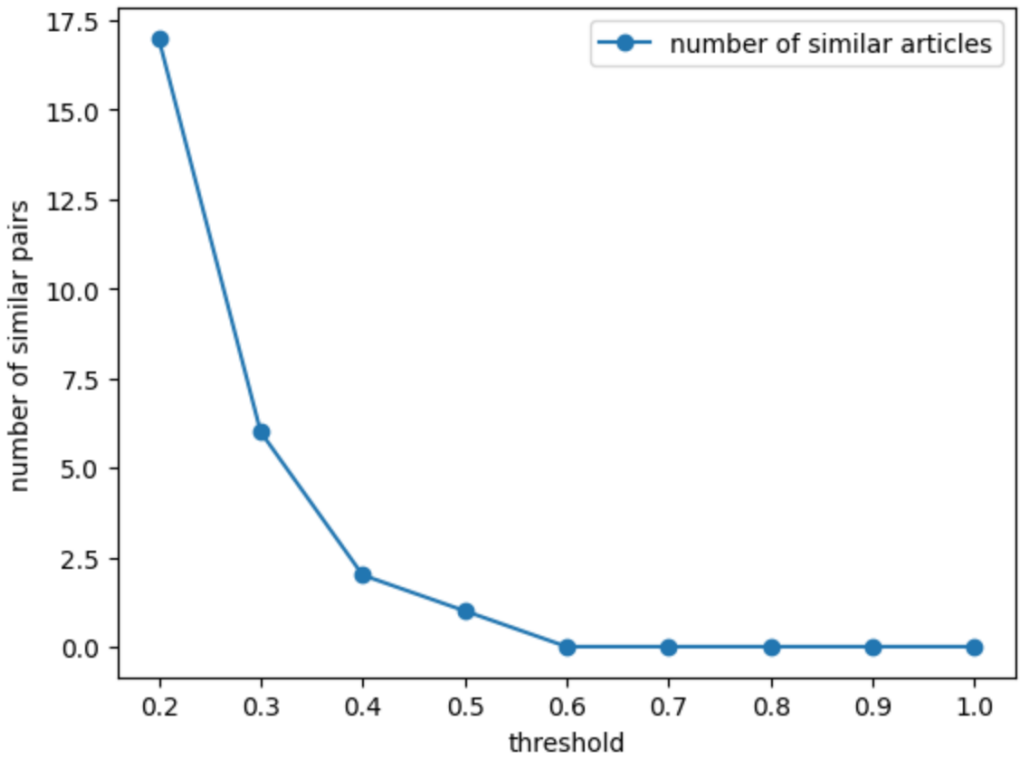
3 The Jaccard Similarity

After the transformation phase the articles are ready to be compared. That is achieved by computing the Jaccard similarity for each pair of different abstracts. Such a measure is very intuitive. Indeed, given two sets, the Jaccard similarity is the ratio between the number of common elements and their overall number. The formula is the following,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \text{ with } J(A, B) \in [0, 1] \tag{1}$$

Then, a threshold on the Jaccard score is defined to discriminate between similar and non similar abstracts. In order to chose an appropriate value, the number of similar articles is investigated for different thresholds as shown in figure 1.

Figure 1: Number of similar articles for different thresholds.



4 Results

Among all 499500 text pairs, the two most similar have a Jaccard score of 0.51, thus meaning they share half of their tokens.

Despite the absence of punctuation makes them difficult to be understood, it is clear that they refer to a common topic. Furthermore, in some cases they share the same sentences. The two detected abstracts are reported hereunder,

[*Reduced coenzyme qcytochrome c reductase from bovine heart mitochondria complex iii was incorporated into phospholipid LDV by the cholate dialysis procedure soybean phospholipids or mixtures of purified phosphatidylcholine phosphatidylethanolamine and cardiolipin could be used oxidation of reduced coenzyme q by the reconstituted vesicles with cytochrome c as oxidant showed the following energycoupling phenomena protons were translocated outward with a coupling ratio he of measurements with mitochondria under similar conditions showed an he ratio of proton translocation was not seen in the presence of uncoupling agents and was in addition to the net acidification of the medium from the overall oxidation reaction potassium ions were taken up by the reconstituted vesicles in the presence of valinomycin in a reaction coupled to electron transfer the coupling ratio for k uptake ke was in the vesicles and approximately in mitochondria the rate of oxidation of reduced coenzyme q by the reconstituted LDV was stimulated up to fold by uncouplers or by valinomycin plus nigericin and k ions addition of valinomycin CT in a k medium caused a transient stimulation of electron transfer the results indicate that SE coupling can be observed with isolated reduced coenzyme qcytochrome c reductase if the enzyme complex is properly incorporated into a phospholipid vesicle.*]

[*Nadhcoenzyme q reductase from bovine heart mitochondria complex i was incorporated into phospholipid LDV by the cholate dialysis procedure mixtures of purified phosphatidylcholine and phosphatidylethanolamine were required oxidation of nadh by coenzyme q catalyzed by the reconstituted vesicles was coupled to proton translocation directed inward with an he ratio greater than similar experiments measuring proton translocation in submitochondrial particles gave an he ratio of the proton translocation in both systems was not seen in the presence of uncoupling agents and was in addition to the net proton uptake from the reduction of coenzyme q by nadh electron transfer in the reconstituted LDV also caused the uptake of the permeant anion tetraphenyl-*

boron the rate of electron transfer by the reconstituted vesicles was stimulated about fold by uncouplers or by valinomycin plus nigericin and k ions the results indicate that energy coupling can be observed with isolated nadhcoenzyme q reductase if the enzyme complex is properly incorporated into a phospholipid vesicle.]

It is true that such a brute-force approach works, leading to correct results, nonetheless it is very time consuming. Indeed, it requires the comparison of a huge number of items, a number destined to increase exponentially with the corpus size. Thereby, a more efficient procedure based on locality-sensitive hashing is proposed as follows.

5 Locality-Sensitive Hashing

Such a technique requires once again shingling, the conversion of documents into sets of elements, in this case, unigrams. Therefore the preprocessing phase is common to both strategies.

One distinction between the two solutions lies in the way these sets are represented and stored. Indeed, the characteristic matrix, which is a very sparse boolean matrix, with columns corresponding to documents and rows to the universal set of tokens, is replaced by a signature matrix, definitely more compact. This latter, built exploiting the MinHash technique, preserves, with a negligible error margin, the Jaccard similarity between the documents.

Another relevant difference of the LSH approach, consists in computing the similarity of the candidate pairs only, which are identified by applying the banding technique to the signature matrix. Since they represent only a small fraction of the entire pairs set, the computational burden is significantly reduced.

The most similar articles are the same obtained with the brute-force approach, but their retrieval required much less time, about one-sixth.

6 Scalability

The analysis is conducted by using PySpark, the Python API for Spark, this last being a data processing framework able to handle massive datasets. It does so by executing the code in a distributed environment, by means of a Java virtual machine. It converts a large dataset in a RDD (Resilient Distributed Dataset), an immutable collection of objects partitionable across a computing cluster and manageable via MapReduce operations. Also, the code is written in Collaboratory, a tool running Jupyter Notebooks, useful for enhancing shareability and exploiting Google servers, thus avoiding to install anything on the local machine. Finally, a special mention to the Spark's Machine Learning library (MLlib), particularly useful for implementing the LSH solution.

7 Conclusions

This project is focused on similarity detection of documents within a very large corpus. Specifically, the analysis is conducted on a fraction of the MeDAL dataset, a huge collection of articles within the biomedical domain. To achieve that objective, documents are first converted into sets of elements and then collected in pairs. For each couple the Jaccard similarity is computed to filter for those most similar, therefore those scoring highest. To speed up computations, an alternative approach based on locality-sensitive hashing is also proposed. Also, to allow for scalability, the code is implemented in PySpark, a distributed framework.

Similarity detection can find different useful applications, among the others, entity resolution and plagiarism recognition. Adapting it for a big data environment, as done in this analysis, can only enlarge its potentials.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.