Document Similarity

May 2020

1. Abstract

Semantic similarity[1] is a metric defined over a set of documents or terms, where the idea of distance between items is based on the likeness of their meaning or semantic content as opposed to lexicographical similarity. These are mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts, or instances, through a numerical description obtained according to the comparison of information supporting their meaning or describing their nature.

Based on text analyses, semantic relatedness between units of language (e.g., words, sentences) can also be estimated using statistical means such as a vector space model[2] to correlate words and textual contexts from a suitable text corpus. The evaluation of the proposed semantic similarity measures is evaluated through two main ways. The former is based on the use of datasets designed by experts and composed of word pairs with semantic similarity/relatedness degree estimation. The second way is based on the integration of the measures inside specific applications such the information retrieval, recommender systems, natural language processing. For example, in natural language processing, knowing one information resource in the internet, it is often of immediate interest to find similar resources.

2. Keywords

Machine learning, Natural Language Processing, Semantic Similarity, Mathematical Tools, Vector Space Model, Normalisation, Cosine Law, Cosine Similarity, Linear Algebra, Inner Product, Information Retrieval

3. Introduction

Plagiarism[3] is the representation of another author's language, thoughts, ideas, or expressions as one's own original work. Plagiarism involves submitting someone's work as their own, taking passages from their own previous work without adding citations, re-writing someone's work without properly citing sources and others.

The basis of a plagiarism checker lies in checking for similar contexts, topics, and word usages in a set of documents. We need a robust algorithm that can maximise on this concept and use words present in a document as features to compare their usage and find the distance or similarity in their context and then convert these similarity scores into plagiarism measures so we can understand how much of the documents share the same content.

We are trying to use concepts of linear algebra – namely vector space models and relationships between them to measure the plagiarism content between two documents. To do this, we need to first pre-process our document and then convert them into document-term matrices so they can be interpreted as a vector space model.

We will be using mathematical vector distance measures such as cosine similarity to find the similarity between the given documents and understand how distance between vectors translates to plagiarism content.

4. Cosine Similarity

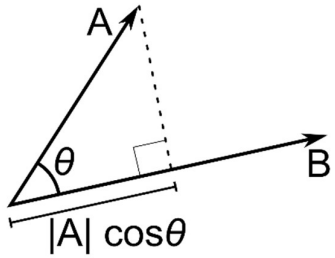
4.1 Linear Algebra Background

Theorem 1. *Given two vectors $\vec{a}$ and $\vec{b}$, the inner product between these vectors is given as $\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + (a_n b_n)$ or $\vec{a} \cdot \vec{b} = a^T b$. Further, the cosine of the angle between these vectors is given as*

$$cos\,\theta = \frac{a^T b}{\|a\|\|b\|} \tag{1}$$

4.2 Proof

Assume two vectors $\vec{A}$ and $\vec{B}$ with angle θ between them and $\vec{p}$ is the projection of $\vec{A}$ on $\vec{B}$.



$$\vec{p} = \vec{A}\,cos\,\theta$$

Taking norm on both sides,

$$\|\vec{p}\| = \left\|\vec{A}\right\| \|cos\,\theta\|$$

$$\vec{p} = k\vec{B}$$

$$k = \frac{B^T A}{B^T B}$$

$$k = \frac{B^T A}{\|B\|^2}$$

Substituting this value of k, we get

$$\vec{p} = \frac{B^T A \vec{B}}{\|B\|^2}$$

$$\|\vec{p}\| = \frac{B^T A}{\|B\|}$$

$$\|\vec{p}\| = \|A\|\,cos\,\theta$$

$$\|A\| \cos\theta = \frac{B^T A}{\|B\|}$$

$$\cos\theta = \frac{B^T A}{\|A\|\|B\|}$$

4.3 Implementation

The input to our problem is a set of 2 documents in raw format. We first perform text pre-processing on these documents to eliminate noise and reduce features and perform dimensionality reduction. We then transform this data into a Bag of Word (BoW) vector space model and further scale these values by performing TF-IDF normalisation with +1 IDF smoothening. We finally use the cosine similarity formula and obtain the distance between these vectors which translates to our percentage of plagiarism content.

5. Algorithm to Perform Cosine Similarity

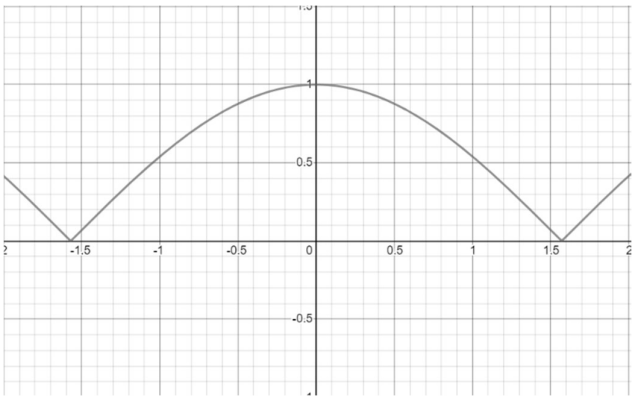
To implement cosine similarity, we find the vector space models of the two documents and then just find the inner product between them.

```
def cosine_similarity(doc1, doc2):
        vectorizer = TfIdf.fit([doc1, doc2], smooth_idf = True)
        vector1 = vectorizer.transform(doc1)
        vector2 = vectorizer.transform(doc2)
        norm = np.linalg.norm(vector1)*np.linalg.norm(vector2)
        return (vector1.transpose()*vector2)/norm
```

While implementing the following algorithm, we ensure to perform

1. Effective text pre-processing to remove lexicons such as symbols and punctuations
2. Remove stop words to reduce noise and dimensions
3. Add +1 smoothening to IDF to reduce chances of 0 division

6. Graphical Analysis



Plot of |cos x|

For any two vector space models, since the magnitude of each dimension is a scaled factor that depends on the term frequency, it will always take values within [0,1] since the term frequency can never be negative. When we apply the inner product between two such vectors, we hence infer that the possible similarities will follow the plot of $|cos\ x|$. We can also verify that if $\theta_1\ is\ lesser\ than\ \theta_2$ in $[-\frac{\pi}{2},\frac{\pi}{2}]$ then $cos\ \theta_1$ *will be greater than* $cos\ \theta_2$. Therefore, lesser the distance, higher the cosine similarity value and higher the plagiarism.

7. Conclusions

On finding the cosine similarity between sets of documents we see that a higher score indicates a higher level of plagiarism and vice versa. Since the cosine similarity measure only uses terms or words as features, it is safe to assume that the checking is performed on a contextual level and not a semantic level. However, since plagiarism majorly depends on the context of the language and according to distributional hypothesis[3] in natural language processing which states that words occurring together have similar meaning, we can say that cosine similarity is an accurate distance metric to find document similarity.

8. Scope of Future Work

Cosine Similarity is a linear algebraic distance measure that is used to find the similarity between two text documents, provided they have been converted into their vector space models. Using more complex word embedding models such as GloVe will help us in storing the semantics as well as context in the document vector models and hence lead to far more accurate scores while estimating the distance between them. We could also use a neural network approach using complex contextual vector models such as Word2Vec, Doc2Vec and even ELMo and BERT to enchance our word embeddings which will give us state of the art distance scores too. Since the algorithm depends vastly on the type of word embeddings used, any model that enhances the vector spaces generated will automatically increase the performance.

9. Summary

- We took a set of 2 documents to find the plagiarism content between them.
- We first performed text pre-processing by removing stop words and unnecessary lexicons.
- We converted these pre-processed documents into word vector models by converting them into TF-IDF vectors.
- We then wrote a function to find the cosine similarity between two vectors
- We passed the vector space models to the above function and found the distance between them.
- Post this we quantified plagiarism mathematically.

References

[1] https://en.wikipedia.org/wiki/Semantic_similarity#Natural_language_processing

[2] https://en.wikipedia.org/wiki/Vector_space_model

[3] https://en.wikipedia.org/wiki/Distributional_semantics#Distributional_hypothesis

[4] R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," 2016 4th International Conference on Cyber and IT Service Management, Bandung, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.

[5] A. Madylova and S. G. Oguducu, "A taxonomy based semantic similarity of documents using the cosine measure," 2009 24th International Symposium on Computer and Information Sciences, Guzelyurt, 2009, pp. 129-134, doi: 10.1109/ISCIS.2009.5291865

[6] Distance Weighted Cosine Similarity Measure for Text Classification by Baoli LiLiping Han

[7] Similarity Measures for Text Document Clustering by Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand

[8] Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering with Shared Nearest Neighbour Method, Lisna Zahrotun