NLP CS5803 IITH Notes

Deepak

March 15, 2024

Contents

1 Introduction

The starting question is how to make computers understand human language. We need to find smart ways of representing the language.

2 Input Representation

Consider text modality, where the input is a document, it consists of words(also called tokens). The document is represented by the set of words/tokens it contains.

Lets see some methods for text representation

2.1 TF-IDF scheme

2.1.1 Formulation

The TF-IDF (Term Frequency-Inverse Document Frequency) scheme is a popular technique used to represent the importance of words in a document corpus.
It combines two factors: term frequency (TF) and inverse document frequency (IDF).
TF measures the frequency of a term in a document. It is calculated by counting the number of occurrences of a term in a document as a raw or by taking a log of it.

$$\mathrm{tf}(t,d) = \begin{cases} 0 & \text{if } c(t,d) = 0 \\ 1 + \log(c(t,d)) & \text{if } c(t,d) \neq 0 \end{cases}$$

where $c(t,d)$ is the count of term t in document d.
IDF de-emphasizes the frequent words across the corpus (all documents combined is usually called corpus) and emphasizes the on words differentiating the documents.

$$\mathrm{idf}(t) = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

where N is the total number of documents in the corpus and df_t is the number of documents containing the term t.
The TF-IDF score for a term in a document is obtained by multiplying its TF value with its IDF value. Mathematically, it can be represented as:

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

Now we get a table with TF-IDF values.

This way, each document is represented by a vector from the column of the table and each word is presented by a vector from the row of the table.

2.1.2 Limitation

- Words are considered at their lexical appearance
- Synonymy is not considered
- Polysemy(word with multiple meanings) is not considered

- long sparse vectors
- context is not considered

2.2 SVD - text representation

2.2.1 Formulation

Consider the term-document matrix and apply SVD to it to do dimensionality reduction, so that we can get dense representations.
Let A be the term-document matrix(matrix with freq. of words in docs), where each row represents a term and each column represents a document, by SVD we have

$$A = U\Sigma V^T$$

U, V are orthonormal matrices and Σ is a diagonal matrix of singular values of A in decreasing order.
By keep only the first k singular values, we have

$$A_k = U_k \Sigma_k V_k^T$$

The k here is much smaller than the original dimension of A. Terms can be represented by the rows of U_k and documents can be represented by the columns of V_k.

3 Results

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

4 Conclusion

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.