Project 2

The goal is to classify documents based on their sentiment

DESCRIPTION

Given a set of documents D containing n documents. For each document d in D, and compute the sentiment s, calculate d_e by applying document embedding (Doc2Vec) to d. Then, classify the documents using the embeddings as features, and the sentiment as labels. Finally, plot the documents, color coded by their sentiment.

EXPECTATION

Documents with similar sentiment should be grouped closely.

TASK BREAKDOWN

1- Pre-process your data for sentiment analysis
2- Train your sentiment analysis algorithm on your pre-processed data
3- Train Doc2Vec on your non-pre-processed data
4- Apply Doc2Vec on each document in your data to form d_{ei} {1 <= i <= n}
5- Compute the sentiment s_i {1 <= i <= n} of each document.
6- Use a classifier (naïve baise, random forest, …) to classify your documents by sentiment
7- Graph the classified documents, color coded by sentiment

TOOLS WITH EXAMPLES

Dataset: Tweets dataset[1]

Document Embedding: Doc2Vec[2]

Really good read on how Doc2Vec works (optional)[3]

Sentiment Analysis + Data preprocessing and cleaning[4]

Sample of classification with Doc2Vec data[5]

Document similarity analysis + visualization[6]

[1] https://raw.githubusercontent.com/kolaveridi/kaggle-Twitter-US-Airline-Sentiment-/master/Tweets.csv
[2] https://radimrehurek.com/gensim/models/doc2vec.html
[3] https://humboldt-wi.github.io/blog/research/information_systems_1718/04topicmodels/
[4] https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/
[5] https://towardsdatascience.com/implementing-multi-class-text-classification-with-doc2vec-df7c3812824d
[6] https://towardsdatascience.com/detecting-document-similarity-with-doc2vec-f8289a9a7db7