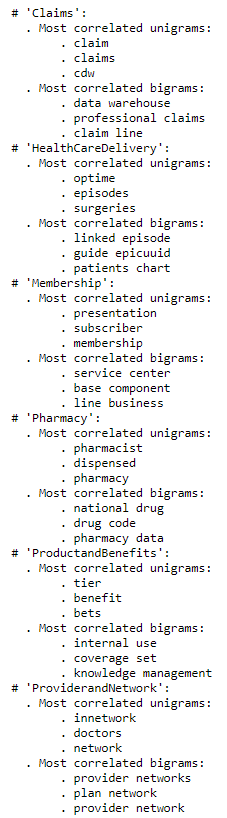
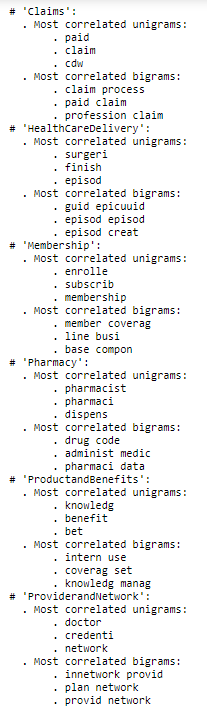


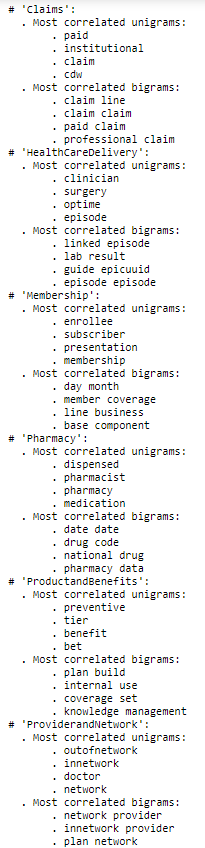
Snowball stemmer

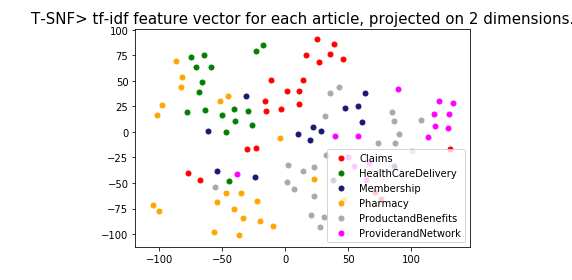


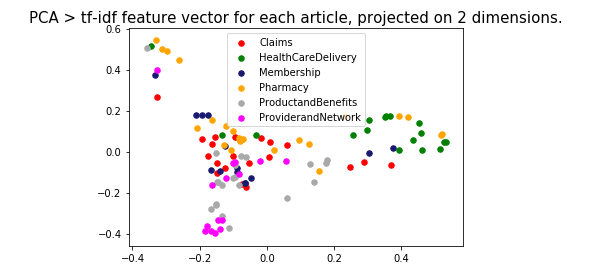
Porter Stemmer

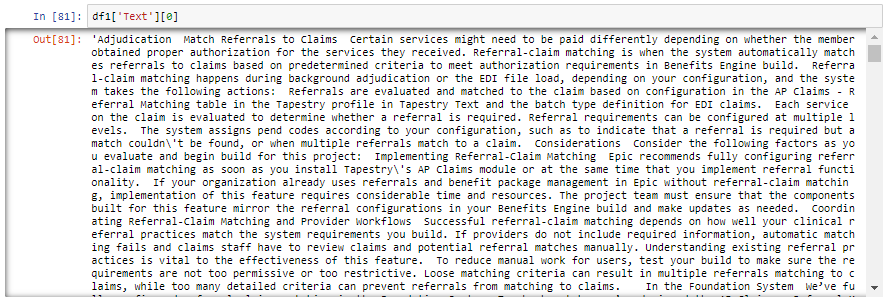


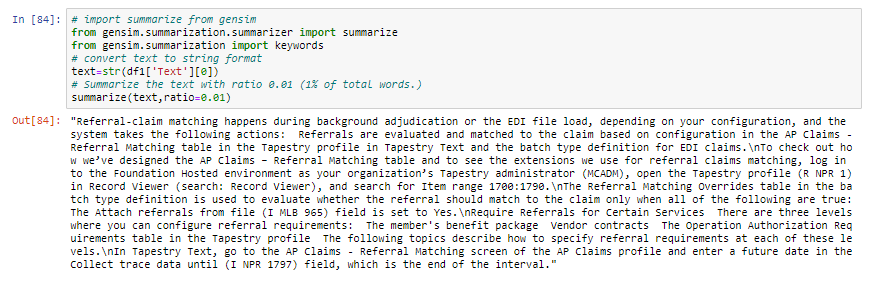
Lemmatize



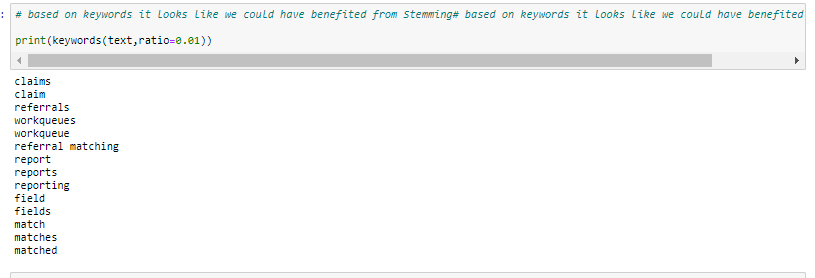








Keywords



Summarization:

Sumy - Library for automatic tect summarization

Methods available:

Luhn - heurestic method

Edmundson heurestic method with previous statistic research

Latent Semantic Analysis, LSA used identify semantically important sentences

LexRank - Unsupervised approach inspired by algorithms PageRank and HITS

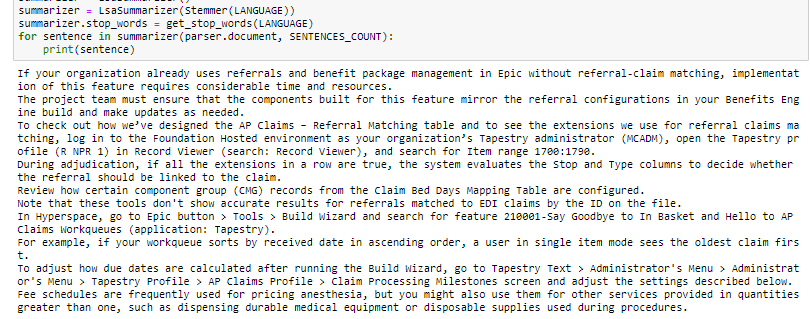
TextRank - Unsupervised approach, also using PageRank algorithm

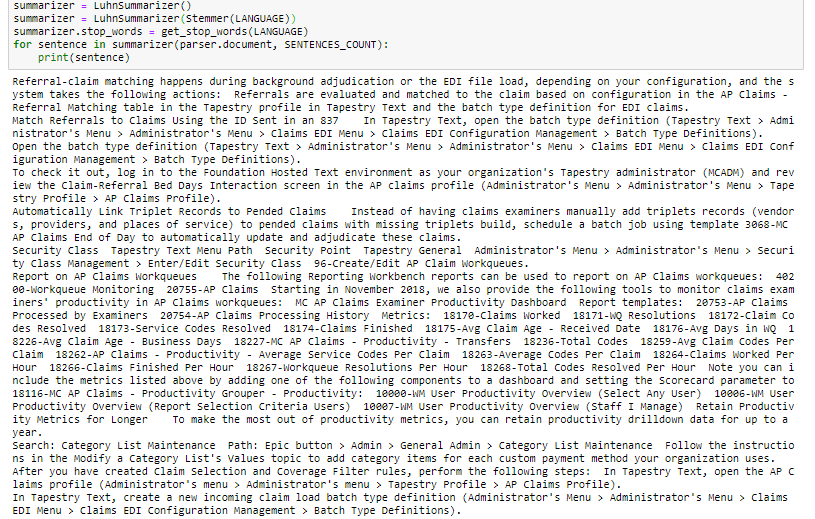
SumBasic - Method that is often used as a baseline in the literature

KL-Sum - Method that greedily adds sentences to a summary so long as it decreases the KL Divergence

Reduction - Graph-based summarization, where a sentence salience is computed as the sum of the weights of its edges to other sentences. The weight of an edge between two sentences is computed in the same manner as TextRank.

Summarization of a 11 page word document



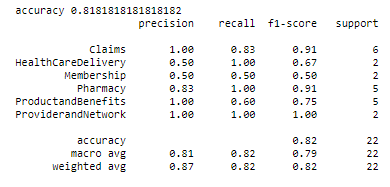


Model 1:

Logistics regression

Word2Vec

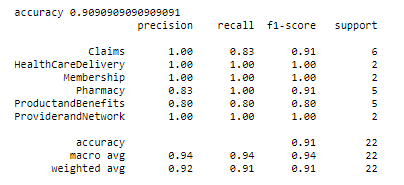
wv = gensim.models.KeyedVectors.load\_word2vec\_format("GoogleNews-vectors-negative300.bin.gz",



Model 2:

Sklearn CountVectorizer()),

('tfidf', TfidfTransformer()),

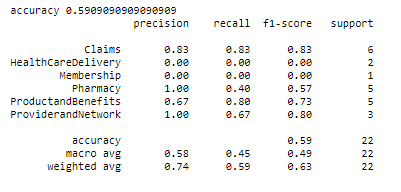


Model 3:

Gensim's Doc2Vec implementation requires each document/paragraph to have a label associated with it.

We do this by using the TaggedDocument method. The format will be "TRAIN\_i" or "TEST\_i" where "i" is

a dummy index of the post.



Reference: <https://github.com/susanli2016/NLP-with-Python/blob/master/Text%20Classification%20model%20selection.ipynb>