Self-Supervised Classifier

Table of contents

Introduction

Self-Supervised Classifier is a set of functions which together comprise a model for classifying sentences as relevant or not. The approach was inspired by Banko et al.'s 2007 "Open Information Extraction from the Web" which used a self-supervised learner to perform open information extraction. We are taking much the same approach to relevancy classification by having the learner tag certain sentences as relevant or irrelevant based on keyword input and then Doc2Vec is trained on these tagged sentences to learn more complex features.

Documentation

convert_to_corpus

- **doc** is the string which you would like to convert to a list of "standardized" words to be part of a corpus.

Sets all of the text to lower case, removes all email addresses, removes all non-alphanumeric characters besides ' and - . All words are lemmatized and then stemmed.

get_TaggedDocuments(pm, instances, iam)

- **pm** is a ProfileManager instance with the information you would like
- **instances** is the number of instances you would like to run
- **iam** is the instance number of the thread you are running [0-instances)

Calls **convert_to_corpus** on the text returned from **pm** and converts them to TaggedDocuments, tagging them as "good" or "bad" if they contain the keywords indicating so.

tag_idks()

Tags all of the "idk_sentences" vectors based on whether they are closer to "good" or "bad".

train_model()

Trains the Doc2Vec model on the TaggedDocuments in "../data/profilemanager/TaggedDocuments"