Automated Document Summarization

Team - Justice League

Machine Learning Term Project

http://github.com/cs60050/ML-JusticeLeague

November 16, 2016

Objective

The Objective is to build an Automatic Text Document Summarization System to help users obtain essence of large documents with least effort using a Machine Learning Approach.

Wikipedia defines Automated Text Summarization as:
Process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document.

DataSet Details

- Coding platform : Python
- Dataset : The CNN News Corpus (500 docs. with summaries)

Enchanted star Amy Adams engaged Now Amy Adams love life is Enchanted,too . The Oscar-nominated Adams is engaged to fellow actor Darren Legallo,her publicist,Cari Ross,confirmed Thursday.Ross did not immediately provide further details . The couple reportedly met in acting class in 2001 , and plan to tie the knot next year. Adams,33,received an Oscar nomination for her supporting role in the 2005 family drama Junebug .She had a breakout role as a Disney-style princess in the 2007 romantic comedy Enchanted .

Amy Adams engaged to actor Darren Legallo, publicist says.Couple met in acting class in 2001, plan to tie the knot next year. Adams, 33, nominated for Oscar for supporting role in Junebug.She also played Disney-style princess in romantic comedy Enchanted

List of Features Used

- TF-ISF
- Text Rank
- Wordnet Ranking
- Numerical Data
- Proper Nouns
- Sentence Positioning
- Sentence Length

Methodology

Classification Function

The summarizer is to be trained with a set of documents and their respective gold standard summaries using supervised learning algorithms. To accomplish this, a classification function that estimates the probability of a sentence being selected in the summary is to be developed.

Feature vector file

A feature vector file was containing feature vectors of all the sentences of the respective training document. A feature vector is a vector of feature values of a sentence.

Two Classes

Feature vectors from all those files are divided into two classes.Those are class of vectors which are taken into summary and class of vectors which are not taken into summary.

Results - Large Summaries

Rouge N-Gram	Avg Precision	Avg Recall	Avg fScore
Rouge-1	0.50600	0.76950	0.59588
Rouge-2	0.46526	0.71521	0.55064
Rouge-3	0.46037	0.71552	0.54740
Rouge-4	0.46004	0.72440	0.54993

Table: Large Summaries

Results - Small Summaries

Rouge N-Gram	Avg Precision	Avg Recall	Avg fScore
Rouge-1	0.53127	0.40905	0.41225
Rouge-2	0.42793	0.33680	0.33869
Rouge-3	0.41708	0.32892	0.33120
Rouge-4	0.41445	0.32786	0.33017

Table: Small Summaries

Thank You