ARSEN KUZMIN

OVERVIEW

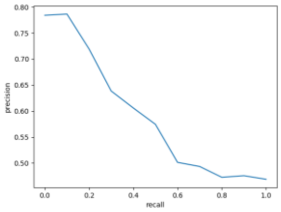
I evaluated local and global methods on cranfield dataset. I evaluated by finding top 30 relevant documents.

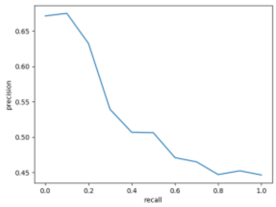
BASE MODEL

Base search engine without any query expansion

Results:

Name of method	Cosine ranking		OKAPI ranking	
	NDSG	Mean average precision	NDSG	Mean average precision
Base	0.457	0.482	0.5376	0.5774





1) Base okapi ranking
2) Base cos ranking

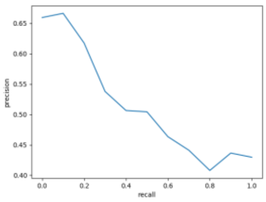
GLOBAL METHODS

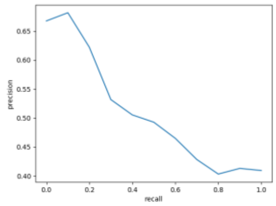
For query expansion I implemented 2 methods:

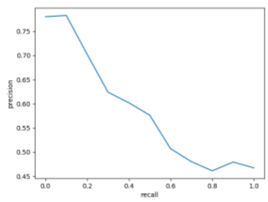
1) Method that uses word2vec model to expand query. Word2Vec is one of technique to learn word embedding using neural network. Word2Vec allow to represent word as a vector so we can define similarity between words. Using similarity, we can find the most similar word for a given word I used library genism to learn word2vec model.
   For each word in query I found the most similar using word2vec and expanded with it query.
   I trained word2vec model on Quora dataset.
2) Method that uses WordNet large lexical database of English. Using WordNet, I expanded query by adding one synonym of each word of query.

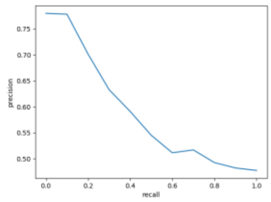
Results:

Name of method	Cosine ranking		OKAPI ranking	
	NDSG	Mean average precision	NDSG	Mean average precision
WordNet	0.4592	0.4781	0.5309	0.5636
Word2Vec	0.4522	0.4889	0.5307	0.5557









1) Word2Vec cosine ranking
2) WordNet cosine ranking
3) Word2Vec OKAPI ranking
4) WordNet OKAPI scoring

LOCAL METHODS

For local query expansion I implemented 2 methods:

1) Rocchio algorithm for relevance feedback. Rocchio algorithm is used to compute new vector with maximum similarity with relevant documents and minimum similarity with nonrelevant documents.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

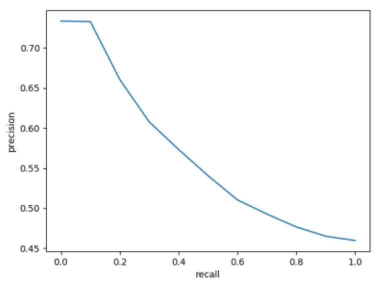
   I used a=1 b=0.75 y =0.15

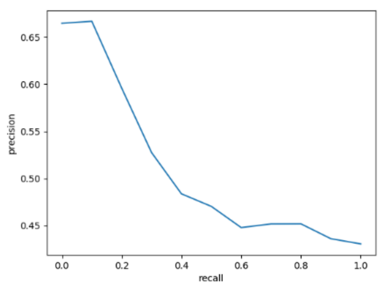
2) Rocchio algorithm for pseudo relevance feedback. In pseudo relevance feedback we assume that the top k ranked documents are relevant, and finally to do Rocchio relevance feedback. I used k = 10

Results:

(There is no meaning to use local methods since OKAP consider only presence of term)

Name of method	Cosine ranking	
	NDSG	Mean average precision
Rocchio algorithm for relevance feedback	0.6529	0.8039
Rocchio algorithm for pseudo relevance feedback.	0.4401	0.4341





1) Rocchio algorithm for relevance feedback
2) Rocchio algorithm for pseudo relevance feedback

Comparison

Rocchio algorithm for relevance feedback algorithm shows much better results than base

Rocchio algorithm for pseudo relevance feedback performs slightly worth than base

DOCUMENT SUMMARIZATION

I used 3 approaches to get summary of document:

1) I extracted 2 first and 2 last sentences of document. In many cases these sentences contain most of information
2) Top n documents that have highest sum of term frequencies of query.
3) Text rank. In text rank we find similarities between sentence vectors are then calculated and stored in a matrix. Then similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation. Finally, a certain number of top-ranked sentences form the final summary

I evaluated these approaches by choosing 2 random texts from this dataset. And used rouge 1 metric. Following result was produced. In second algorithm text I used: 'museum' as query for first text and 'architecture' for second text.

```
2 first and last sentences
Real summary: Orhan Pamuk: Monumental state treasure-houses such as the Louvre or the Met ignore the stories of the individual. Exhibitions should become ever more intimate and local
Summary I love museums and I am not alone in finding that they make me happier with each passing day.
I take museums very seriously, and that sometimes leads me to angry, forceful thoughts.
11 The future of museums is inside our own homes.
12 The picture is, in fact, simple.
___________________________
rogue 1 score: 0.0416666666666666

Query dependent
Real summary: Orhan Pamuk: Monumental state treasure-houses such as the Louvre or the Met ignore the stories of the individual. Exhibitions should become ever more intimate and local
Summary sign up to the art weekly email

that is not to understate the importance of the louvre, metropolitan museum, topkapд± palace, british museum, prado, and pinacoteca вђ" all of which are veritable treasures of humankind.5 the measure of a museum's success should not be its ability to represen
___________________________
rogue 1 score: 0.10101010101010101

Text rank
Real summary: Orhan Pamuk: Monumental state treasure-houses such as the Louvre or the Met ignore the stories of the individual. Exhibitions should become ever more intimate and local
Summary National museums, then, should be like novels; but they are not.
The aim of big, state-sponsored museums, on the other hand, is to represent the state.
3 We are sick and tired of museums that try to construct historical narratives of a society, community, team, nation, state, people, company or species.


___________________________
rogue 1 score: 0.1

2 first and last sentences
Real summary: Modernism produced perhaps the biggest variety of styles in history, from concrete wigwams and penguin pools to streets in the sky, says Steve Rose
Summary The nautical theme has been reduced to a corny joke in British seaside architecture, but there's a dignified restraint to it here.
The modernist tides of 1930s Europe washed this elegant culture palace up on our shores thanks to an enlightened patron (Earl De La Warr, mayor of Bexhill) and two Г©migrГ© architects (German Eric Mendelsohn and Chechen Serge Chermayeff).
It is concrete and glass and little else.
The massive, mushroom-headed columns enabled large, practical spans and gave the building a dynamic, sculptural quality, while natural light was brought in through the all-glass elevations and light wells.
___________________________
rogue 1 score: 0.14492753623188406

Query dependent
Real summary: Modernism produced perhaps the biggest variety of styles in history, from concrete wigwams and penguin pools to streets in the sky, says Steve Rose
Summary the nautical theme has been reduced to a corny joke in british seaside architecture, but there's a dignified restraint to it here.9. arnos grove underground station, north london
this 1932 building is proof that modern architecture could achieve a civic presence even within the historic patchwork of a city like london, though sadly the neighbourhood built around it never matched the station's clarity of form and intent.10. boots facto
with a bluntness that betrays his roots in engineering rather than architecture, owen williams laid out boots' 1932 nottingham manufacturing complex to a functionalist design.
___________________________
rogue 1 score: 0.1363636363636363638
```

For the first text

1st approach works quite well. 3rd sentence match with the real summary.

2nd approach performs works better than first one because 2nd sentence match better.

3rd approach is the best from these 1st and 3rd sentences display real summary.

For the second

1st approach shows good results since 2nd and 3rd sentences describe main concept of the article

2nd approach performs not well on this text. Quite hard to understand main concept.

3rd performs not well too since Modernism has not mentioned

References:

Text rank:

https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

Word2Vec:

https://radimrehurek.com/gensim/models/word2vec.html

Rogue score

https://github.com/bdusell/rougescore/blob/master/rougescore/rougescore.py

README:

In order to launch code following libraries should be installed:

genism 3.7.2

networkx 2.3

nltk 3.4

matplotlib 3.0.3

smart-open 1.8.1

scipy 1.2.1

numpy 1.16.2

Word2Vec

In order to use word2vec you can download ready model from link or train by yourself by downloading quora dataset: link and changing path_to_dataset variable in word2vec_expansion.py file.

Evaluation

Change PATH_TO_DATA variable in main.py to path where cranfield dataset located.

Part1

To reproduce results of part1 run main.py

Part2

To reproduce results of part1 run doc_summary_evaluation.py

If you want to plot results change variable in evaluation.py file eleven_points_interpolated_avg function plot=True.