11-693 Software Methods for Biotechnology Homework1 Report

Shen Wu

Error Analysis

qid	rel	Similarity	Rank	Text
1	99			Classical music is dying
1	0	0.2673	2	Pop music has absorbed influences from most other genres of popular music
1	1	0.4522	1	Classical music may never be the most popular music
1	0	0.1768	3	Everybody knows classical music when they hear it
2	99			Energy plays an important role in climate change
2	0	0	2	Old wine and friends improve with age
2	0	0	2	With clothes the new are the best, with friends the old are the best
2	1	0.0981	1	Climate change and energy use are two sides of the same coin
3	99			One's best friend is oneself
3	1	0.4629	1	The best mirror is an old friend
3	0	0.3961	2	My best friend is the one who brings out the best in me
3	0	0.1667	3	The best antiques are old friends
MRR		1		

Although this looks like perfect result because MRR == 1, but it's not. Because some lucky mistakes is made. For example, not all the words were normalized, therefore "One" can't match to "one".

To avoid this mistakes, I normalize every token before storing them. The result is as following:

qid	rel	Similarity	Rank	Text
1	99			Classical music is dying
1	0	0.2673	3	Pop music has absorbed influences from most other genres of popular music
1	1	0.4522	1	Classical music may never be the most popular music
1	0	0.3536	2	Everybody knows classical music when they hear it
2	99			Energy plays an important role in climate change
2	0	0	2	Old wine and friends improve with age
2	0	0	2	With clothes the new are the best, with friends the old are the best
2	1	0.2941	1	Climate change and energy use are two sides of the same coin
3	99			One's best friend is oneself
3	1	0.4629	2	The best mirror is an old friend
3	0	0.4950	1	My best friend is the one who brings out the best in me
3	0	0.1667	3	The best antiques are old friends
MRR		0.83333		

This time, MRR reduced to 0.833. One problem of the system is that many meaningless words are involved. For example: "The", "a', and so on.

To further improve the system, I filtered out all the meaningless words, i.e. stop words.

The experiment result is as following:

qid	rel	Similarity	Rank	Text
1	99			Classical music is dying
1	0	0.3849	3	Pop music has absorbed influences from most other genres of popular music
1	1	0.61237	1	Classical music may never be the most popular music
1	0	0.516397	2	Everybody knows classical music when they hear it
2	99			Energy plays an important role in climate change
2	0	0	2	Old wine and friends improve with age
2	0	0	2	With clothes the new are the best, with friends the old are the best
2	1	0.43301	1	Climate change and energy use are two sides of the same coin
3	99			One's best friend is oneself
3	1	0.44721	2	The best mirror is an old friend
3	0	0.7071067	1	My best friend is the one who brings out the best in me
3	0	0.2236067	3	The best antiques are old friends
MRR		0.83333		

If looking at MRR, this time the performance is not improved at all. But if looking at every score, the confidence of right answer is greatly improved, which means the system can give right answer with much higher confidence now.

Further improvement:
I changed the way of computing cosin similarity: instead of building whole vector according to the overall dictionary, I set the dictionary to be all the worlds in standard sentence. That makes the calculation a little faster also reduced the impact of unused words.
I also tried with stemming. Although theoretically the performance should improved, but it didn't, it dropped. I believe that's because the scale of data is too small.
If given a larger data set, I would train the weight of each word, which hopefully will further improve performance.