RECOMMENDER SYSTEMS

Report submission as a requirement for the module of "Information Retrieval Systems"

MSc Degree in Data Analysis
at the school of computing
Robert Gordon University
Aberdeen, Scotland
December 2020

Question 2a)

Lexic

Lexic analyses the structure of words, It includes processeses such as Tokenization, stemming, stop word removal and others. These methods simplifies the structure of the text and it then creates a bag of words that can be used to compare word similarity with the ones on a user query.

For an example on how Lexical techqniques work I will be using the phrase below. I have used Rapid Miner and the text transformation techniques mentioned above to get a lexical focus representation of the sentence.

"Information retrieval systems is my favorite module in all the world"

inform retriev system favorit modul world

Figure 1 Rapid Miner lexical text transform techniques

Syntactic

Syntactic methods tend to grammatically label the words in a phrase "usually after this has been lexically structured"; by doing so it creates relations between the words of a phrase allowing for a computer to make sense of it and better interpreting it.

Below is an example on how our sample sentence can be syntactically analised. The example has been provided thanks to "Google NLP demo API"

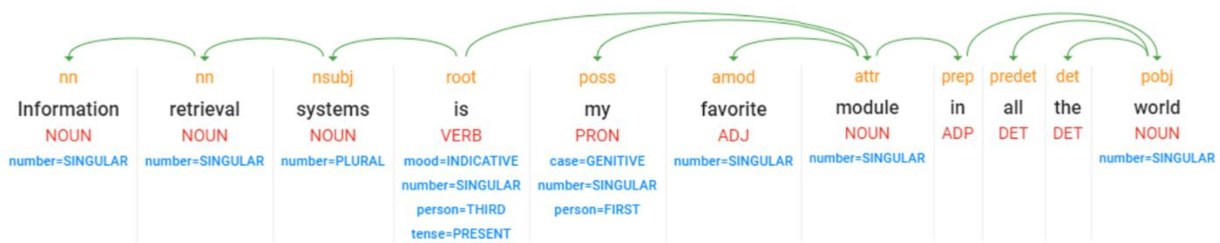

Figure 2 Syntactic analysis for the sample sentence

Semantic

This method requires understanding the meaning and intention conveyed in a user query. In order to do so there are many techniques used in order to add context to a phrase " usually after having the [hrases Syntactically analysied".

Below we can see examples of some of these techniques such as sentiment analysis or name entity recognition which categorizes word or words in a sentence as people, place, person, quantity, company or location. Below the word world has been tagged as a location, and module as a consumer good which makes sense because I am the consumer which is undertaking a module in IR at RGU.

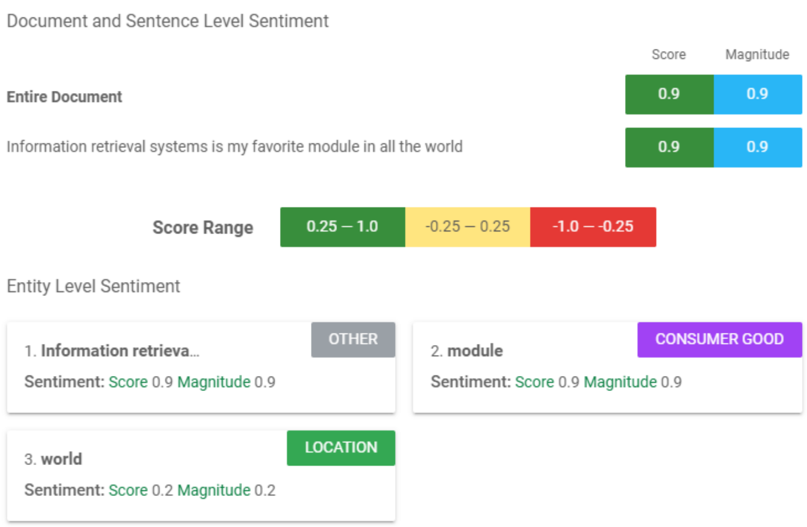

Figure 3 Semantic analysis of sample sentence

Ignoring syntactic and semantic analysis and using only lexical analysis will create problems with ambiguity (e.g. Synonymy and Polysemy) this means that multiple words and phrases might have the same or different meanings depending on the context in which they are used. This problem can lead with mismatch of vocabulary in a document or query. In order to close this intent gap and semantic gap and disambiguate the keyword search we can use different approaches that allow the implementation of syntactic and semantic solutions. See some of these approaches below.

Taxonomy

It is known as the art of naming. It focuses on describing and categorizing things or concepts. Like in the semantic example we saw above this useful technique allows for the creation of an entity tree like structure which nodes are characterized by labels as the ones seen in the example below. These tree structures are then used by search engines to match the keywords from the user query to the ones in the indexed documents.

Figure 4 Taxonomy example

Feedback

This technique focuses on obtaining information from the user in order to optimize the search engine system. It can involves the user judgment to weather a page is relevant, it can also be implicit feedback which assumes that documents which the user has clicked on are found to be important. Some search engines also tried to ask the users to rate the accuracy of the search results. This feedback is then implemented by the Rocchio technique which moves the user query closer towards the documents that were found more relevant by other users.

Word Embeddings

This technique tries to examine how often certain word or phrase "A" appears close to word or phrase "B". This creates a relation between concepts in the data. In other words it creates a bag of words which are related to each other. These high-dimensional vectors capture the meaning of words. The idea is that words that appear in similar context will be considered close and their respective vectors will contain similar values, this way words such as cat, lion, kitty, tiger will be associated by a set of values. The words cat and dog can also be related by the word pet. It basically creates a human like memory where all words and concepts are linked between each other depending on the environment that person was raised for the machine this means the data that it was used to create and train the model. The previous approcahes mentioned before can help with the word association and wight change and manipulation for a more accurate vector space model.

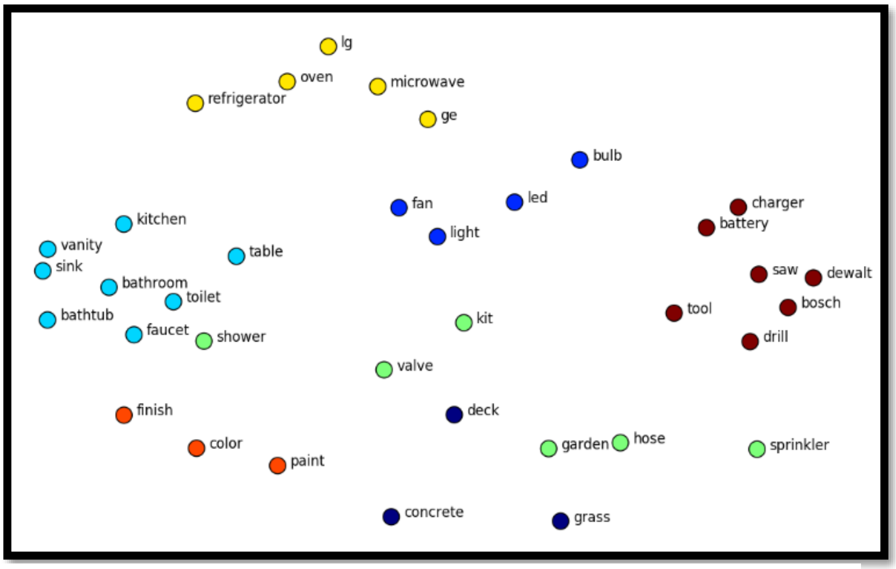

Figure 5 Example of word embedding

Question 2b)

Advantages

Task 2 discussed in our coursework do not contain feedback from many users which means that a collabortive recommender system would not be effective. In contrast we have a user model which works effectively with content base recommender systems, we also have content base information with the "title" and "description" for each of the products. It allows us to inmediately recommend content and avoid a cold start and sparcity which would be the case if using a collaborative system. If new entried where introduces we would be able to recommend them instantly. It also allows us to recommend items without these being popular as well as avoiding popular bias. It allows us to provide explanation for the recommendations.

Disadvantages

We have to pre-process and structure the data in a way that allows us to use it efficiently so that sentiment and intention can be captured to enrich the data at a granular leve. We can’t use quality judgment from other users as is just content based. We also need to create a model for every new entry so that it can be used by the recommender system.

Question 3b)

For this question I am going to set up changes in model and parameters to try to find the one tha yields the best results. My initial setup is an Item-base KNN by 3 nearest neighbours using the pearson coefficient. I will attempt to tweak the model, the number of K and test for better results. Later I will discuss and compare the results.

Prediction Model	Correlation mode	# K	RMSE	MAE
User KNN	pearson	3	1.034	0.799
User KNN	pearson	5,8,13	1.032	0.797
User KNN	cosine	3,5,8,13	1.040	0.811
Item KNN	pearson	3	0.790	0.600
Item KNN	Pearson	8	0.771	0.583
Item KNN	pearson	13	0.767	0.581
Item KNN	cosine	3,5,8,13	1. +	0.7 +

Prediction Model	# Factors	LR	Iterations	RMSE	MAE
Matrix Factorization	10	0.5	30	0.198	0.088
Matrix Factorization	21	0.05	30	0.190	0.075
Matrix Factorization	34	0.05	30	0.187	0.074

The results show when using “user-base KNN” the results where inferior to the one in “Item-base KNN”. It looks like computing similarities scores between users did not blend well with the project, this is perhaps because there are more variables that need to be considered when building user recommender systems such as bought/viewed etc… There are much more items than there are users which is why the item-base KNN perfoemed better. I also ran different numbers for K and coefficients and got the same result for both models. Seems like cosine performs the worst compared to pearson. It also seems to show a better result when increasing the number of K. Increasing the number of K means that we are using more users to obtaine the average rating for each item hence obtain a more robust answer. It is important not to increase the number of K as to make the solution inneficient and slow when the difference in MAE is not too relevant.

I also used a third model " Matrix Factorization" this model was the one used for the " Netflix Prize" for Netflix's recommender system. It benefits itself froma very simple dataset avoiding noise which can greatly decrease the effectiveness of this solution. It is also the lightest solution (i.e. decreased sparity) because it simplifies the matrix with the million fo customers and items making it lighter. It work best with small datasets. The results in MAE for this algorithm proved to be the best achieving an 0.074 MAE a hughe difference compared to the best solution obtained by our KNN. It seems that for recommender systems a facotrized matrix allows the relation between interactions of users and items to really show up.

Question 3b)

Mean Absolute Error or MAE is a quantity used to measure how close predictions are to the eventual outcome. It informs how bad a recommender system is by measuring the distance between the difference in prediction and the actual outcome in other words the error, and then computes the average of all the errors in the dataset and returns the result; The lower the result the smaller the error hence the better the algorithm. The only draw back is that it fails to penalize large errors. This means that if our recommender system rating system is not 1 to 1 and it grows exponentially from one rating to another it won't take this fact into consideration. There are other Evaluation techniques such as Root Mean Squarred Error or RMSE which also measures the average magnitude of the error meaning that our calculation si sentitive to large errors. RMSE is a more popular metric mainly because of its reliability when handling the variance of the frequency distribution of error magnitudes.

For a news article recommender system it would be useful to include evaluation criterias such as diversity, tho that it covers user different interests, we don't want to keep the user in a bubble of the same information, there are many diverse things that happen around the world and traying to offer some of this news can help keep a user aware of what is happening in the world. Similar to diversity is serendipity which tries to also shows unexpected but interesting news happening around the world. Another very important criteria would be User demographics could also play an important role as different races, ages and sec might be interested in different articles. Recommender persistence for articles we know the user would be very interested but perhaps hasn't had the time to view them, this would avoid the user in missing out important news. All of the previous metrics. All of the previous mentioned measured can be applied together with A/B testing, this statistical test helps understanding user engagement and satisfaction from online real time features such as new news articles. There are many big companies using this measurement (e.g. LinkedIn, Instagram and even Facebook) allowing for them to understand their customers and providing a better user experience.