


Al-Ghossein

For each received document, we evaluate the topic model and process the document to update the underlying model. We use the first 20% documents of the stream to train the model which is evaluated on the remaining documents.

The goal in document modeling is to maximize the likelihood on unseen documents (D_{test}), given a trained topic model. Perplexity measures the ability of a model to generalize to new data and is used to evaluate topic models

 image-20220528202818122

A lower value of perplexity indicates a better generalization capacity

 image-20220528203014814

. The red dotted vertical line marks the detection of a drift by AWILDA

Then, for online recommendation, it compares using $recall@N$ and $DCG@N$ (usado para comparar ranking)


 image-20220528215155779

Muni

given a new ticket, rank all historical tickets by similarity. The ones with similarity score above a given threshold (0.8) were chosen

The resulting ranking was:

1. Compared with an existing automated ranking, that was some semantic similarity (SS) rule based analysis
2. Manually evaluated


 image-20220528204723678

We randomly pick 10% of each data set (tickets) as the test set and 20% of the data set as the training set

Sund

The end results of the project were evaluated together with Vizrt to determine if using a word embedding model was viable for their needs.

Five query documents of relatively recent dates were randomly selected. These were considered as our test set. For each document in the test set, each model would retrieve the n most similar documents. For each model, the five query documents and their n most similar documents, including their respective similarity score, were written to disk for Nils [the domain specialist] to perform manual evaluation.

 image-20220528205922185

“prone to bias and error, due to having a single person perform the evaluation.”

Wahba

Their data was imbalanced; They considered doing oversampling and undersampling, but decided not to

They compared the different embedding models by added a downstream classifier (for the category of the ticket) and evaluating the classifier performance, with standard metrics of Precision, Recall and F-measure

Zhou

Given a ticket, the model recommended the solution

The text of the solution was compared via jaccard similarity with the ground thruth text of the solution

If the similarity were above a threshold, it was considered a hit

Weighted Accuracy was measured

They also used a similar metric, recommending many solutions are taking the average of the accuracy (by the same method above)

Here we define a recommended resolution as a hit if it has a jaccard similarity greater than a threshold with the ground truth resolution.

Nathália (TCC)

estudo com os usuários,
armazenando as informações sobre as ações dos mesmos através do feedback explícito

avaliações do usuário através de feedback explícito é através da interação dos usuários com os botões de “gostei” e “não gostei”

As avaliações feitas pelos usuários também serão avaliadas através da métrica de Precision@k sendo k igual à 10 e 5. A métrica Recall que normalmente acompanha a métrica anterior, não será avaliada pois para isso, seria necessário conhecer o total de itens relevantes recomendados

 image-20220528214852321

questionário com perguntas sobre a experiência do usuário no sistema tais como as opiniões sobre as recomendações e opiniões sobre o sistema.

DeLucia

As a pseudo-baseline we included the results from an Elasticsearch “more like this” query and a “naive” calculation of percent words in common.

Sentence bert paper

sentence pairs of STS benchmarks

linkeding guys (detext)

online evaluation: CTR@5

offline: NDCG@10

their online evaluation is very similar to what I want to do (but I'll do it offline)

Resumo

Todos os métodos usaram um dos seguintes métodos:

1. comparam com algum modelo existente: eu nem tenho um modelo existente e nem gostaria de assumir que algum outro método é o “correto”
2. fez uma comparou a similaridade textual (quão parecidas as palavras eram): me parece uma forma tola de avaliar, visto que me parece muito relevante que o modelo seja capaz de recomendar ticket que sejam parecidos no seu assunto, mais do que nas palavras utilizadas.
3. Fazendo uma avaliação “online” com usuários usando o sistema: não me é viável para comparar várias técnicas, sendo que algumas delas podem ser bem ruins (e eu não posso sujeitar usuários reais à elas)
4. Utilizou um conjunto muito pequeno de testes
5. Utilizou alguma métrica não supervisionada, tipo “perplexidade”, que até agora eu não consegui entender
6. usou SENTECE PAIRS

Related works in representation based document ranking

Ranking Relevance in Yahoo Search

<https://www.semanticscholar.org/paper/Ranking-Relevance-in-Yahoo-Search-Yin-Hu/60fbb8dc34bbff762a7d6a3b5e70c00ebe0ccd0b>

representation based

Towards Deep and Representation Learning for Talent Search at LinkedIn

<https://arxiv.org/abs/1809.06473>

representation based

Learning deep structured semantic models for web search using clickthrough data

<https://paperswithcode.com/paper/learning-deep-structured-semantic-models-for>

Representation based