Université de Genève

Information Retrieval
14x011

TP 3 : Zip's Law and Precision/Recall

Author: Sajaendra Thevamanoharan

E-mail: Sajaendra.Thevamanoharan@etu.unige.ch

Author: Deniz Sungurtekin

E-mail: Deniz.Sungurtekin@etu.unige.ch

18 April 2020

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Introduction

This lab will be divided into two main parts. The first part is an evaluation of the search engine that we built in our previous labs. The second part is to study the statistics of word occurences in texts using Zipf's Law.

Evaluation of the search Engine

We plot the PR curves for a search Engine in several settings to analyse the quality of retrieval. Precision (P) will assess the density of relevant documents when exploring the ranked list downwards. Recall (R) indicates the proportion of relevant documents retrived so far.

At one step n of such a scrolling of the ranked list one may place the threshold 'positive above, negative below' and compute the related P_n and R_n as follows :

$$P_n = \frac{T_p}{T_p + F_p} \tag{1}$$

$$R_n = \frac{T_p}{T_p + F_n} \tag{2}$$

Precision is defined as the number of true positives over the number of true positives plus the number of false positives. Recall is defined as the number of true positives over the number of true positives plus the number of false negatives.

True positives is the number of documents that are considered as relevant. False positive is the number of documents that are considered relevant but not relevant in the reality. False negative is the number of documents that are considered non relavant but relavant in the reality.

1. We create a function that takes as input the ranked, the relevant list, n and providing the corresponding (R_n, P_n).

2. We took N = 15, corresponding to the number of texts in the nasa corpus. We create two queries. Using the function defined just before, we plot the PR curve for the given two queries.

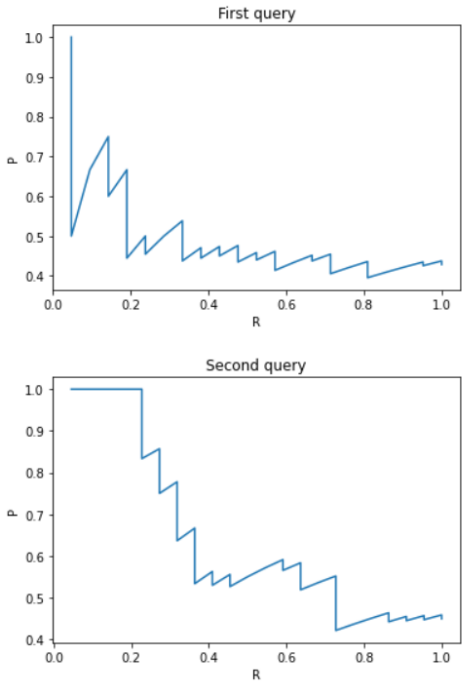

Figure 1: PR curve for the given two queries

3. We average these two P-R curves using the 11-point technique. In the 11-pt interpolated average precision, we are looking at 11 recall levels (0.0, 0.1, 0.2, ..., 1.0) and finding the interpolated precision at each point. We average these scores across all of the different queries or information needs to find our system's score.

We have for these two curves the following average curve.

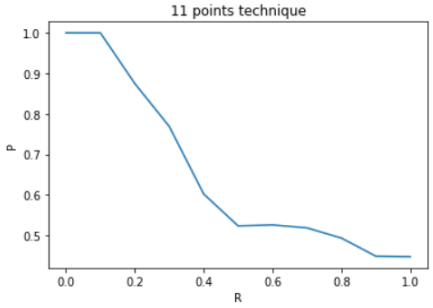

Figure 2: Average of two previous PR curves using 11-point technique

The typical usage in information retrieval is for evaluation. It helps us to compare one system to another, but where it really helps is in letting us compare how your system is changing as you tweak parameters.

We can notice that the system deliver a decent performance with a good quality of retrieval.

Verify the Zipf 's Law

For this section, we will choose a French book as our corpus to verify Zipf's Law. This Law tells us that a small number of words occur very frequently and that many words occur rarely. Moreover, it states that the probability of encoutering the r-th most common word is inversely proportional to its rank r:

$$P(r) = 0.1/r \tag{3}$$

So, we choose a book named "20000 Lieues Sous Les Mers", compute the frequency of each word and arranges them in a list according to their rank to verify if it follows the Zipf's Law. On the same diagram, we will plot the words distribution, the corresponding Zipf's Law and a linear regression line of our distribution to see if it can be used to explore the dependence between word's probabilities and their ranks:

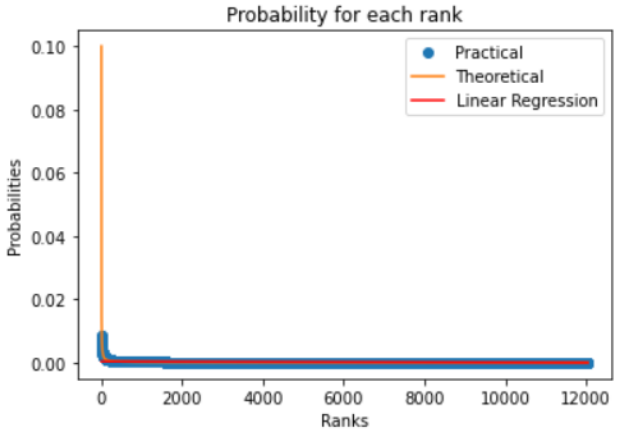

Figure 3: Probabilities over rank

We can clearly see that the distribution of our book follow the Zipf's Law but in a different scale. Effectively, we observe that a lot of words are very rare but the most frequent words are less frequent than in the Zipf's Law (Which is explained by the fact that we removed the stop words). However, the linear regression can't be used to explore the dependence between word's probabilities and their ranks because its not enoughly accurate to estimate a hyperbolic line. The result gives use a straight line near 0 which is logical because most of the words occur rarely.

Furthermore, here are the R-squared and p values:

```
R-squared: 0.113162
P-Value: 0.000000
```

Figure 4: Observed value

The R-squared tell us how well the regression model fits the observed data, so clearly our R-squared is not good as explained because its very hard to estimate a hyperbolic line with only a linear regression. The P-value, is a number describing how likely it is that our data would have occurred by random chance. Moreover, our P-value is less than 0.05 so it is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct. Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

Here, we will observe the residual plot for the linear regression:

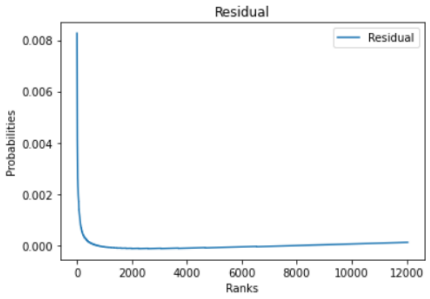

Figure 5: Residual

We almost obtain the distribution of our book which is logical because our regression is a line near 0 so when we compute the difference between the original data and the regression we obtain the same original data with litlle changes.

Now, we will show 10 examples of extremely frequent, very rare, and averagely frequent words:

```
4 Very frequent words:  nemo capitain _nautilus_ mer
3 Averagely Frequent Words:  être dit heur
3 Very Rare Words:  chéri chère châteaux
```

Figure 6: Some words

The book is about sea, so we have words like "Nemo", "capitaine" and "mer" that have high frequency which is coherent. Obviously, this categories will be the more useful in information retrieval if we take off stop words and words having a high inverse frequency. In this example, we take off those words so this categorie is the more useful in information retrieval. However, if we don't do this operation the very frequent words might be those stop words which have also high inverse frequency and the more useful words in information retrieval will be the averagely frequent words.