Assignment 4

Fall 2016
CS834 Introduction to Information Retrieval
Dr. Michael Nelson

Mathew Chaney

December 15, 2016

Contents

List of Figures

List of Tables

1 Question 10.5

1.1 Question

Find a community-based question answering site on the Web and ask two questions, one that is low-quality and one that is high-quality. Describe the answer quality of each question.

1.2 Approach

Yahoo! Answers is the question and answering website I chose to answer this question. First, the high quality question I asked was "How is the probability of precipitation calculated by meteorologists?".

I deem this question to be of high quality based on grammaticality, spelling and punctuation, as well as the use of concise terminology that may be slightly obscure to an uneducated reader.

It is also a focused question that has a definite set of answers. There are a finite number of ways to perform the calculation in question, and they are well known. The ideal answerer is even included as part of the question. Refer to Table 1 to see some example answers for this question.

Question: How is the probability of precipitation calculated by meteorologists?
Trump is Putin's *****
Ye
we got hakerzed in the ***
no
Penis

Table 1: Question and responses of good quality.

For the poor quality question I asked "Trump is not russia, y u no belief he got hakerz to do it to us?". It begins with a premise that doesn't make sense, even the most uneducated person wouldn't suspect that Trump *is* Russia. It also lacks focus as to what is being asked or and has no real context and is full of misspellings and terrible grammar. Correspondingly, the answers were lacking content, using poor grammar and exhibiting numerous spelling errors, and some possibly offensive to some readers. Refer to Table 2 for the question again with some sample answers.

Question: Trump is not russia, y u no belief he got hakerz to do it to us?
Trump is Putin's *****
Ye
we got hakerzed in the ***
no
Penis

Table 2: Question and responses of poor quality.

2 Question 10.6

2.1 Question

Find two examples of document filtering systems on the Web. How do they build a profile for your information need? Is the system static or adaptive?

2.2 Answer

The first example I found was www.amazon.com. This is an Internet marketplace where nearly any type of product can be purchased online and delivered to one's home. The document filtering Amazon performs is done in an adaptive manner using a combination of the user data from previous purchases to recommend new items the user may be interested in. The system can also recommend items that similar users often purchase that are also similar to items the user is viewing while browsing the item repository. For example, if one purchases a mechanical pencil it is likely that the user will also be interested in purchasing graphite refill packages or extra replacement erasers. This system changes over time as the database of related items is based on the past purchases of the different items and how some items are more often purchased together than others.

3 Question 11.2

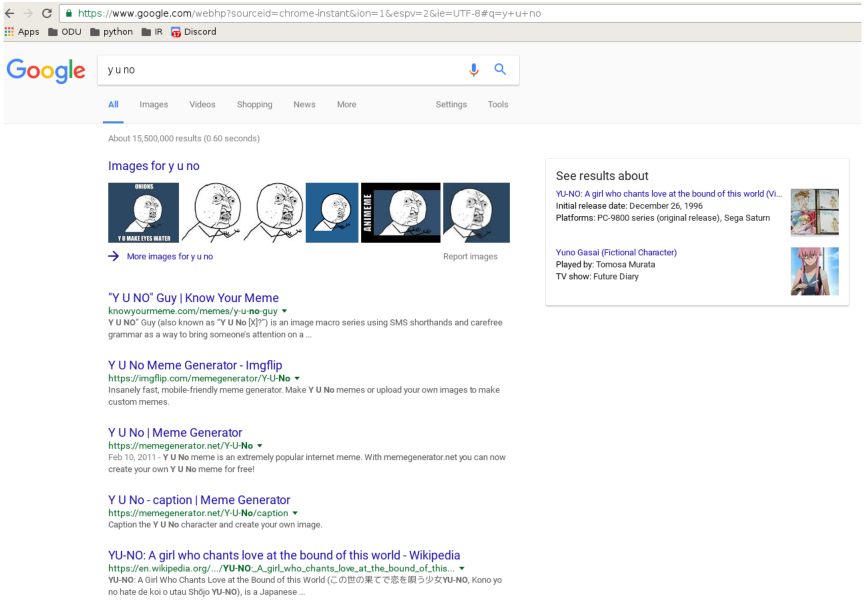
3.1 Question

Does your favorite web search engine use a bag of words representation? How can you tell whether it does or doesn't?

3.2 Answer

My favorite search engine is Google. I am fairly certain they do not limit their retrieval model to the simple bag-of-words model described by the text book.

As an example, there is a well known online meme, or trend, called "y u no guy". It is an image macro used to bring attention to something and uses a popular practice for online forums of shortening words to their root sound corresponding to a letter, e.g. “you” becomes “u”. The target message is accompanied by a cartoon man with a face of frustration and rage. See Figure 1 for the results of issuing the query “y u no” to Google's search engine.



About 15,500,000 results (0.60 seconds)

Images for y u no

"Y U NO" Guy | Know Your Meme
knowyourmeme.com/memes/y-u-**no**-guy
Y U NO" Guy (also known as "**Y U No** [X]?") is an image macro series using SMS shorthands and carefree grammar as a way to bring someone's attention on a ...

Y U No Meme Generator - Imgflip
https://imgflip.com/memegenerator/Y-U-**No**
Insanely fast, mobile-friendly meme generator. Make **Y U No** memes or upload your own images to make custom memes.

Y U No | Meme Generator
https://memegenerator.net/Y-U-**No**
Feb 10, 2011 - **Y U No** meme is an extremely popular internet meme. With memegenerator.net you can now create your own **Y U No** meme for free!

Y U No - caption | Meme Generator
https://memegenerator.net/Y-U-**No**/caption
Caption the **Y U No** character and create your own image.

YU-NO: A girl who chants love at the bound of this world - Wikipedia
https://en.wikipedia.org/.../**YU-NO**:_A_girl_who_chants_love_at_the_bound_of_this...
YU-NO: A Girl Who Chants Love at the Bound of this World (この世の果てで恋を唄う少女**YU-NO**, Kono yo no hate de koi o utau Shōjo **YU-NO**), is a Japanese ...

See results about

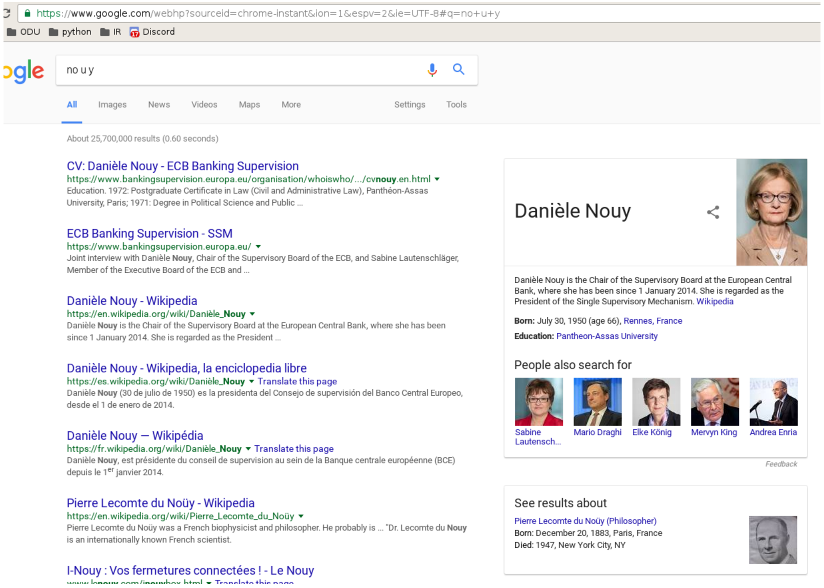
YU-NO: A girl who chants love at the bound of this world (Vi...
Initial release date: December 26, 1996
Platforms: PC-9800 series (original release), Sega Saturn

Yuno Gasai (Fictional Character)
Played by: Tomosa Murata
TV show: Future Diary

Figure 1: y u no guy

Google recognizes these three simple terms to be part of a phrase and retrieves results based on that phrase with a high degree of accuracy.

To show that Google's engine does not use the bag-of-words model, consider this similar search with the terms "no u y".


https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=no+u+y
ODU
python
IR
Discord
no u y
All
Images
News
Videos
Maps
More
Settings
Tools
Feedback

About 25,700,000 results (0.60 seconds)

CV: Danièle Nouy - ECB Banking Supervision
https://www.bankingsupervision.europa.eu/organisation/whoiswho/.../cv**nouy**.en.html
Education. 1972: Postgraduate Certificate in Law (Civil and Administrative Law), Panthéon-Assas University, Paris; 1971: Degree in Political Science and Public ...

ECB Banking Supervision - SSM
https://www.bankingsupervision.europa.eu/
Joint interview with Danièle **Nouy**, Chair of the Supervisory Board of the ECB, and Sabine Lautenschläger, Member of the Executive Board of the ECB and ...

Danièle Nouy - Wikipedia
https://en.wikipedia.org/wiki/Danièle_**Nouy**
Danièle **Nouy** is the Chair of the Supervisory Board at the European Central Bank, where she has been since 1 January 2014. She is regarded as the President ...

Danièle Nouy - Wikipedia, la enciclopedia libre
https://es.wikipedia.org/wiki/Danièle_**Nouy** Translate this page
Danièle **Nouy** (30 de julio de 1950) es la presidenta del Consejo de supervisión del Banco Central Europeo, desde el 1 de enero de 2014.

Danièle Nouy — Wikipédia
https://fr.wikipedia.org/wiki/Danièle_**Nouy** Translate this page
Danièle **Nouy**, est présidente du conseil de supervision au sein de la Banque centrale européenne (BCE) depuis le 1er janvier 2014.

Pierre Lecomte du Noüy - Wikipedia
https://en.wikipedia.org/wiki/Pierre_Lecomte_du_Noüy
Pierre Lecomte du Noüy was a French biophysicist and philosopher. He probably is ... "Dr. Lecomte du **Nouy** is an internationally known French scientist.

I-Nouy : Vos fermetures connectées ! - Le Nouy

Danièle Nouy

Danièle Nouy is the Chair of the Supervisory Board at the European Central Bank, where she has been since 1 January 2014. She is regarded as the President of the Single Supervisory Mechanism. Wikipedia

Born: July 30, 1950 (age 66), Rennes, France

Education: Pantheon-Assas University

People also search for

Sabine Lautensch... Mario Draghi Elke König Mervyn King Andrea Enria

See results about

Pierre Lecomte du Noüy (Philosopher)
Born: December 20, 1883, Paris, France
Died: 1947, New York City, NY

Figure 2: Daniele Nouy

With the same three terms in a different order, a completely different ranking is presented by the search engine. The results for this search are not regarding y u no guy at all, they are dominated by pages about the Chair of the Supervisory Board at the European Central Bank, Daniele Nouy. This shows that Google's engine accounts for query term ordering, which would not be present in a plain bag-of-words retrieval model.

4 Question 11.4

4.1 Question

Show how the linear feature-based ranking function is related to the abstract ranking model from Chapter 5.

4.2 Answer

Starting with the first mention of the abstract ranking model, refer to Figure 3. This is an example of the ranking function for a single document.

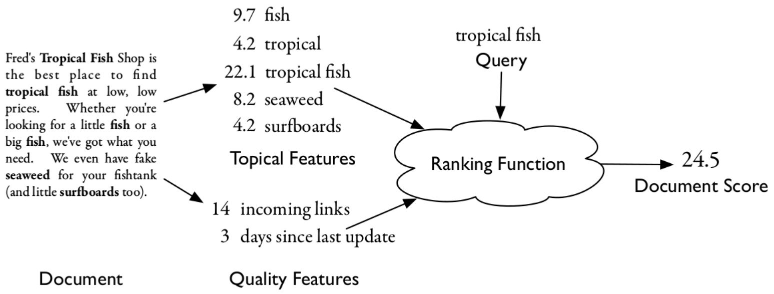

Fred's Tropical Fish Shop is the best place to find tropical fish at low, low prices. Whether you're looking for a little fish or a big fish, we've got what you need. We even have fake seaweed for your fishtank (and little surfboards too).
9.7 fish
4.2 tropical
22.1 tropical fish
8.2 seaweed
4.2 surfboards
Topical Features
tropical fish
Query
Ranking Function
24.5
Document Score
14 incoming links
3 days since last update
Document
Quality Features

Figure 3: The components of the abstract model of ranking...

And a more formal definition of the abstract ranking model can be found in equation 1:

$$R(Q, D) = \sum_i g_i(Q) f_i(D) \tag{1}$$

This is a linear combination of two feature functions, f_i is a feature function that extracts a score from the document and g_i is a feature function that extracts a score from the query.

Now, for a definition of the linear feature-based retrieval model, refer to Equation 2:

$$S_\Lambda(D; Q) = \sum_j \lambda_j \cdot f_j(D, Q) + Z \tag{2}$$

here, f_j is a feature function that maps query/document pairs to real values, i.e. scores, so it is also a linear combination of functions that emit scores based on features of some related piece of the components of the formulation, either the document, the query, some parameter ($\lambda \in \Lambda$), or a constant that is not related to the document but could be related to the query (Z). This is very similar to how the abstract ranking model functions in that it is a summation of a group of scoring functions over the elements to be ranked. The mechanism is the same as before with the abstract model, with the addition of the λ parameters.

5 SVM Light Example

5.1 Inductive Example

SVM Light [1] and the Inductive example were downloaded from http://www.cs.cornell.edu/People/tj/svm_light/.

Following the instructions in the Inductive example, the `svm_learn` program was invoked on the training data file train.dat, the output from which can be found in Listing 1. This produced the model file that is used in the classification step.

```
[mchaney@mchaney-l svmlight]$ ./svm_learn train.dat model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..
900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..
OK. (2000 examples read)
Setting default regularization parameter C=1.0000
Optimizing.......................................................
................................................................
................................................................
................................................................
................................................................
................................................................
........done. (425 iterations)
Optimization finished (5 misclassified, maxdiff=0.00085).
Runtime in cpu-seconds: 0.07
Number of SV: 878 (including 117 at upper bound)
L1 loss: loss=35.67674
Norm of weight vector: |w|=19.55576
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=383.42791
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=5.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>95.40% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>93.07% (rho=1.00,depth=0)
Number of kernel evaluations: 45954
Writing model file...done
```

Listing 1: output of the svm_learn program using the train.dat data file

After the model file was generated the example directs the user to run the `svm_classify` program on the provided test data `(filename: test.dat)` using the training model `(filename: model)` from the previous step. The output of this can be found in Listing 2.

```
[mchaney@mchaney-l svmlight]$ ./svm_classify test.dat model predictions
Reading model...OK. (878 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 97.67% (586 correct, 14 incorrect, 600 total)
Precision/recall on test set: 96.43%/99.00%
```

Listing 2: output of the svm_classify program using the model training data file

5.2 Discussion

As the results show, the precision was 96.67% and the recall was 99.00%, which are very high scores for these measures.

6 References

[1] Thorsten Joachims. Svm light. Available at: http://www.cs.cornell.edu/People/tj/svm_light/. Accessed: 2016/12/13.