SET Project Report, Fall 2012

Clustering Web Search Results (Software Project)

Rohan Agrawal

UNI: ra2616

Motivation:

Sometimes web search queries are ambiguous and it is impossible for web search engines to decipher the meaning intended by the user. For, e.g. If a user types in the query apple, it is unclear whether he wants search results related to Apple, the company or Apple, the fruit. Some sort of feedback is needed from the user, to make the search query less ambiguous. Instead of having the user type out in detail, what he is looking for, it would be useful if he could see the various interpretations of his search term, allowing him to expand his search in the intended direction through a single click. This will make the users searching experience less tedious. Grouping similar type of results is a clustering problem, which I have based my software project on.

Tools/ Data:

I have used

- Bing API to get web search results (which are clustered)
- NLTK: used porter stemmer
- pyYaml: NLTK depends on pyYaml to run.
- Scikits-Learn: I have used the k-means++ algorithm implemented by this package.
- pyDocs: To generate documentation.

System Description:

There are 4 broad steps:

1. Data preprocessing
2. Finding a suitable value of k – the number of clusters
3. Clustering
4. Cluster label generation or Post Processing

1. Data Preprocessing

First data is retrieved from the Bing search API. Data consist of web site URL, page title and Short snippet. All of the text from the website is not used, as this would considerably slow down the generation of results, which is unwanted. This text is tokenized, stemmed through Porter Stemmer (NLTK implementation) and stop words are removed (the list of stop words is taken from Salton and Buckley, http://www.lextek.com/manuals/onix/stopwords2.html). A few web specific words such as html, jsp and aspx are added to the list of stop words. URL's are processed in a slightly different way. The text after the third '/' character is used for tokenization, as the part before this contains no valuable keywords for cluster labeling or even clusters generation. Sites like ebay and amazon contain a lot of noisy character in their URL's; hence words are not extracted from their URLs. Also, words that appear in more than 80% of the documents are deemed as corpus specific stop words and hence removed. Since they appear so frequently, they cannot influence the output of the clustering algorithm and will not make good cluster labels at all.

Once all of the data from the web search results has been preprocessed, periods are replaced by a specific sentence separator, to avoid generating phrases from the data which may span over 2 sentences. Such phrases with a sentence separator in between them are ignored as candidate cluster labels, described later in Cluster Label Generation.

The whole data set is then converted into its vector space representation where a word is represented by its tf-idf score. The 200 words with highest tf-idf score are taken to be the features for k-means clustering. The data set is converted in to an n * m sparse array with n = Number of documents, m = number of features.

2. Finding a suitable value of k – the number of clusters

The main drawback of k-means algorithm is that it has to be given the optimal number of clusters. In [1], Weiss has directly used the optimal value of k in the Descriptive k-means algorithm. This cannot be done here, and hence the number of clusters k is found using a Cost Function.

$$Cost\ C = Inertia\ (of a\ clusering\ with\ k\ centers) + k * Penalty$$

The penalty is taken to be 1.15 based on empirical experiments. As mentioned in most of the literature on Cluster Labeling, it is a very subjective task; therefore I relied on my judgment to decide the optimal value of the penalty function. There is a tradeoff in setting a high penalty values versus a low penalty value. A high penalty value would prevent further clustering and the user could be left without a cluster he/she desired. Thus, I have set this variable on the lower side. I have varied the value of k from 1 to 10, and stopped when the value of C increases. The k value corresponding to the lowest C value is chosen to be the optimal value of k.

3. Clustering

K-means++ algorithm described in [2] is used for clustering, due to its speed advantage over Hierarchical Clustering ($O(n^3)\ vs\ O(IKMN)$, where I is the number of maximum iterations the algorithm is allowed to run, K is the number of clusters, N is the number of rows of data and M is the number of features).

The difference between k-means and k-means++ is the initialization scheme of k-means++ is designed to make it converge faster than k-means, thus helping our purpose. Initialization is done in a way, such that cluster centers are chosen using a weighted probability distribution such that the probability of a point being chosen as a cluster is directly proportional to the square of distances between the point and existing cluster centers.

The remaining 2 steps: re-assignment of vectors to centroids, and re-computation of centroids is done in the same way. The distance between 2 vectors is given by the sum of squares of the tf-idf values of their features.

Since K-Means++ is a randomized algorithm, I have done 100 trials of the algorithm and then selected the trail run that has the lowest value of Cost C, described in (3. Custering).

4. Cluster label generation or Post Processing

Cluster labeling is the biggest challenge and research area in Web search result clustering. In [1] Weiss has used a descriptive clustering approach, in which features are selected to be good descriptors of a label. Nouns are preferred as features. For extracting nouns, text chunking has to be done, and this will be an extra burden on the system in terms of processing time. Another approach has been to use Wikipedia as a source for generating a cluster name, based on the high weighted words in the centroids as described in [3]. The drawback of this scheme is there could be certain clusters which are not represented by a Wikipedia page. For e.g. while clustering the search term "polo" , the Bing search API returns a couple of results based on Polo Resources Limited, but this company does not have a Wikipedia page and hence cannot be represented by a cluster label taken from Wikipedia.

In my approach, I have used the highest weighted phrases according to the tf-idf weighing scheme. Each of the top 20 phrases by tf-idf value in each cluster centroid has been considered as a candidate cluster label. Since words/phrases appearing in more than 80% of the documents are removed as domain specific stop words, there are not a lot of phrases in the candidate cluster label set, that contain common words, or words that don't represent their cluster. But they do contain a lot of repetition, for e.g. while testing my system on the search term "polo", the candidate cluster labels were,

'polo men', 'polo men women', 'polo women', and so on

Thus labels which did not represent any extra information over other labels were removed, i.e. "polo men" and "polo women" were removed due to their terms being a subset of the terms contained in "polo men women" label. Labels that were a reordering of the search terms are also removed, for e.g. for the search term "snow leopard", the label "leopard snow" will be removed. After this pruning process, the labels which had the highest 3 weights are all considered as representative cluster labels, and are show to the user. In the software system made for this project, the user can click on any label to expand his search. The search term is automatically concatenated with the chosen label and the user gets new results based on his selected cluster, which can be clustered even further, if the user so desires.

Evaluation:

There are 2 parts of the evaluation for this project. First, do the clusters have high purity, i.e. documents are grouped within their true clusters. Second, evaluation of the cluster labels which is purely subjective.

Tests are carried out on 3 search terms,

1. Polo
2. Jaguar
3. Apple

I have stored the search results and cluster labels for these 3 test runs in the ra2616-SETproject/SETProjectDemo/ directory in 3 separate files called polo.txt, jaguar.txt and apple.txt. Calculation of Purity and evaluation of correctness of cluster labels is described in more details in those files.

1. Polo

All cluster labels accurately describe their respective clusters. The Purity is 95% which is very high (19 / 20 documents correctly clustered).

2. Jaguar

2 / 3 cluster labels accurately describe their respective clusters and the remaining cluster label also gives its cluster a satisfactory label. The Purity is 85% which is also high (17 / 20 documents correctly clustered).

3. Apple

5 / 6 cluster labels accurately describe their respective clusters. The remaining cluster is a bit ambiguous as it deals with 2 separate topics. The Purity is 85% which is high (17 / 20 documents correctly clustered).

Details about the above results can be seen in the above mentioned 3 files.

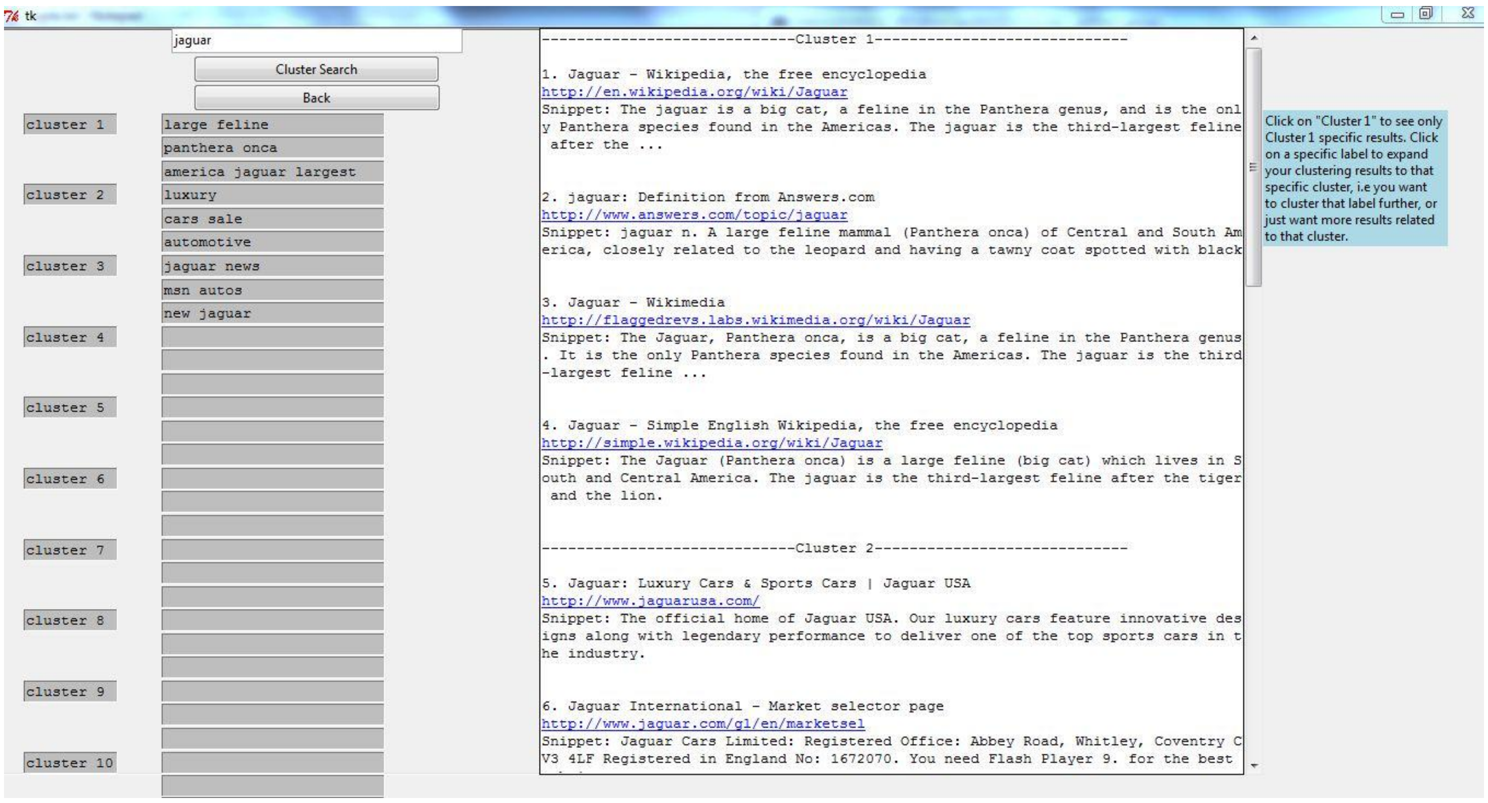
How does my system perform the task I set out to do?

My system does what I had mentioned in the project proposal, except for the fact that I wanted to make a graph visualization of the clusters which would show nodes as web sites, edges between nodes would have represented similarity between documents.

Improvements:

As mentioned above a graphical interface of the web search results can be used to make the users search experience richer. I would have also liked to put up my project on the web, but could not do so because my main focus was on the clustering and labeling part.

Demo Output:

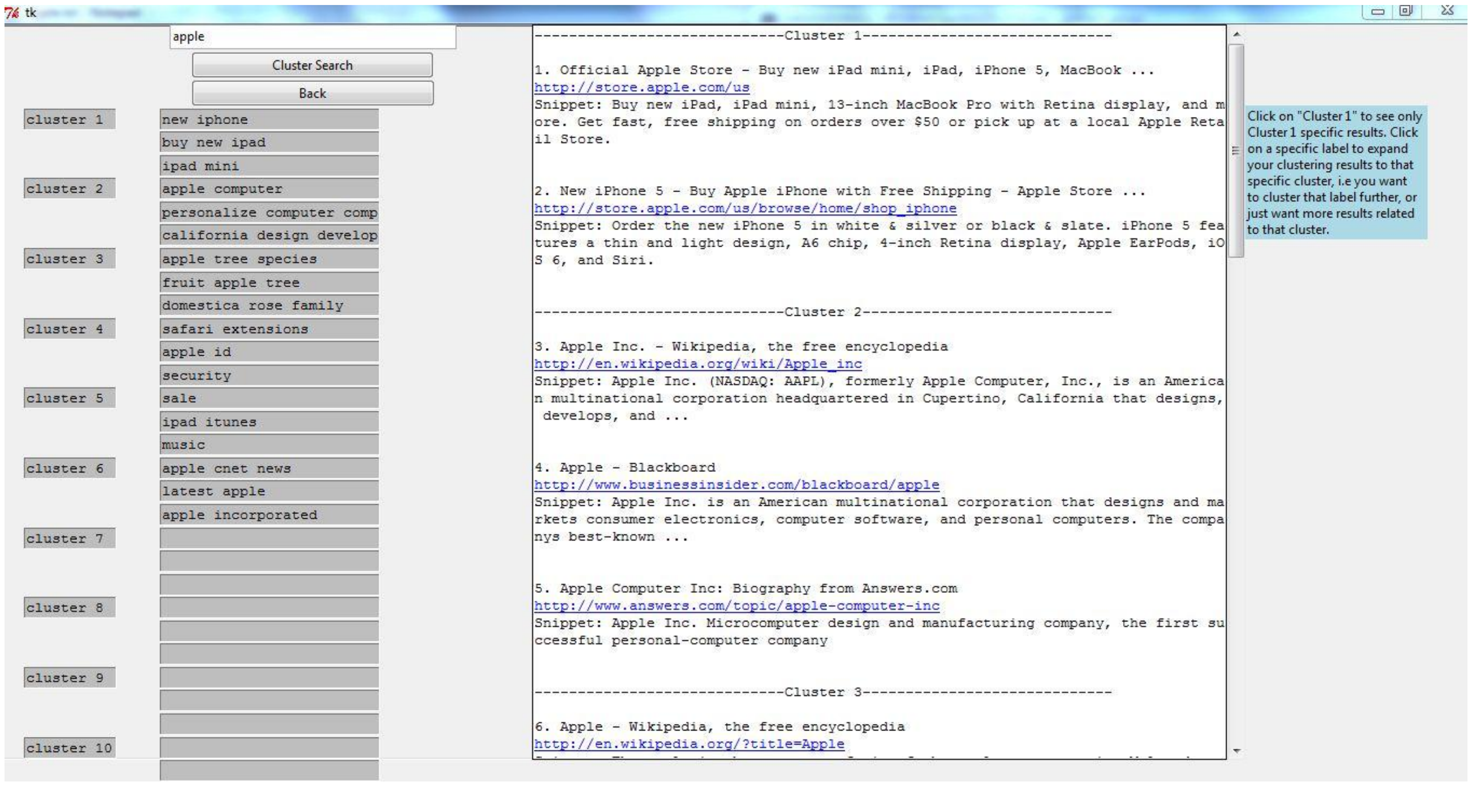


```
cluster 1    large feline
             panthera onca
             america jaguar largest
cluster 2    luxury
             cars sale
             automotive
cluster 3    jaguar news
             msn autos
             new jaguar
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
cluster 9
cluster 10
```

```
-----------------------------Cluster 1-----------------------------

1. Jaguar - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Jaguar
Snippet: The jaguar is a big cat, a feline in the Panthera genus, and is the onl
y Panthera species found in the Americas. The jaguar is the third-largest feline
 after the ...


2. jaguar: Definition from Answers.com
http://www.answers.com/topic/jaguar
Snippet: jaguar n. A large feline mammal (Panthera onca) of Central and South Am
erica, closely related to the leopard and having a tawny coat spotted with black


3. Jaguar - Wikimedia
http://flaggedrevs.labs.wikimedia.org/wiki/Jaguar
Snippet: The Jaguar, Panthera onca, is a big cat, a feline in the Panthera genus
. It is the only Panthera species found in the Americas. The jaguar is the third
-largest feline ...


4. Jaguar - Simple English Wikipedia, the free encyclopedia
http://simple.wikipedia.org/wiki/Jaguar
Snippet: The Jaguar (Panthera onca) is a large feline (big cat) which lives in S
outh and Central America. The jaguar is the third-largest feline after the tiger
 and the lion.


-----------------------------Cluster 2-----------------------------

5. Jaguar: Luxury Cars & Sports Cars | Jaguar USA
http://www.jaguarusa.com/
Snippet: The official home of Jaguar USA. Our luxury cars feature innovative des
igns along with legendary performance to deliver one of the top sports cars in t
he industry.


6. Jaguar International - Market selector page
http://www.jaguar.com/gl/en/marketsel
Snippet: Jaguar Cars Limited: Registered Office: Abbey Road, Whitley, Coventry C
V3 4LF Registered in England No: 1672070. You need Flash Player 9. for the best
```

Click on "Cluster 1" to see only Cluster 1 specific results. Click on a specific label to expand your clustering results to that specific cluster, i.e you want to cluster that label further, or just want more results related to that cluster.

Figure 1: Search results for the search term jaguar



```
cluster 1    new iphone
             buy new ipad
             ipad mini
cluster 2    apple computer
             personalize computer comp
             california design develop
cluster 3    apple tree species
             fruit apple tree
             domestica rose family
cluster 4    safari extensions
             apple id
             security
cluster 5    sale
             ipad itunes
             music
cluster 6    apple cnet news
             latest apple
             apple incorporated
cluster 7
cluster 8
cluster 9
cluster 10
```

```
----------------------------Cluster 1----------------------------

1. Official Apple Store - Buy new iPad mini, iPad, iPhone 5, MacBook ...
http://store.apple.com/us
Snippet: Buy new iPad, iPad mini, 13-inch MacBook Pro with Retina display, and m
ore. Get fast, free shipping on orders over $50 or pick up at a local Apple Reta
il Store.


2. New iPhone 5 - Buy Apple iPhone with Free Shipping - Apple Store ...
http://store.apple.com/us/browse/home/shop_iphone
Snippet: Order the new iPhone 5 in white & silver or black & slate. iPhone 5 fea
tures a thin and light design, A6 chip, 4-inch Retina display, Apple EarPods, iO
S 6, and Siri.


----------------------------Cluster 2----------------------------

3. Apple Inc. - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Apple_inc
Snippet: Apple Inc. (NASDAQ: AAPL), formerly Apple Computer, Inc., is an America
n multinational corporation headquartered in Cupertino, California that designs,
 develops, and ...


4. Apple - Blackboard
http://www.businessinsider.com/blackboard/apple
Snippet: Apple Inc. is an American multinational corporation that designs and ma
rkets consumer electronics, computer software, and personal computers. The compa
nys best-known ...


5. Apple Computer Inc: Biography from Answers.com
http://www.answers.com/topic/apple-computer-inc
Snippet: Apple Inc. Microcomputer design and manufacturing company, the first su
ccessful personal-computer company


----------------------------Cluster 3----------------------------

6. Apple - Wikipedia, the free encyclopedia
http://en.wikipedia.org/?title=Apple
```

Click on "Cluster1" to see only Cluster1 specific results. Click on a specific label to expand your clustering results to that specific cluster, i.e you want to cluster that label further, or just want more results related to that cluster.

Figure 2: Search results for the search term apple

I have stored the results of 3 test searches 'apple'; 'jaguar' and 'polo' in the ra2616-SETproject/SETProjectDemo/ directory in 3 separate files.

References:

[1] Weiss, D. (2006). Descriptive clustering as a method for exploring text collections (Doctoral dissertation, University of Technology).

[2] Arthur, D. and Vassilvitskii, S. (2007). "*k*-means++: the advantages of careful seeding". *Proceedings of the* eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035.

[3] Carmel, D., Roitman, H., & Zwerdling, N. (2009, July). Enhancing cluster labeling using wikipedia. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 139-146). ACM.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1). Cambridge: Cambridge University Press.