Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 4: Programming Project
November 5, 2018
Due dates:
Stage 1 - November 20, 2018; Tuesday: Class Time
Stage 2 - December 11, 2018; Tuesday: Class Time
Stage 3 - December 27, 2018; Thursday: Class Time ~~December 25, 2018; Tuesday: Class Time~~

Date of Last Update: December 12, 2018; 3:55 pm

This is a programming type assignment. You have to work with your term project partner(s). For different purposes or stages you may use different toolkits or programming languages. Make sure that you start early. From each group only one person, group leader of the term project, will submit the required items as defined below. The groups are the same as term project groups, see the end of this document.

Purpose: Clustering experiments by using a test collection with incremental growth. Comparing the effectiveness of two clustering algorithms in terms of producing cohesive clusters. In your experiments use the first 1000 documents of the Bilkent Information Retrieval Milliyet test collection (see github). For obtaining the clusters first you will use the first 100 documents of the collection, obtain clusters and measure the cohesion of clusters under different conditions as defined below. Repeat the same for first 200, 300, ..., 1000 documents; with increments of 100 documents.

It provides 1. Experience with the state of the art information retrieval toolkits, 2. Experimental design for statistical tests, and 4. Experience of writing a scientific paper.

Stage Definitions

Stage 1: Involves Step 1 defined below.
Indexing results: Give statistical information about each collection of size 100, 200, ... 1000. Possible items to include, but not limited by, are the following: The change in number of terms as you index more documents, change in indexing file sizes, statistical observations about stopwords such as their percentage in the texts. Only a hardcopy of the report is required; please bring your report to class.

Stage 2: Steps 2-6
Clustering Results & Statistical Test: The report involves the experimental results.
Only hardcopy is required; please bring your report to class.

Stage 3: Step 7
Paper: In the form of a ACM conference paper.
Both hardcopy and pdf copy are required bring it to class and send a pdf file for the paper and with it send a rar file for the programs. In the subject line write: CS533: Programming Project Material

Steps

1. Use the first 1000 documents of the Bilkent Information Retrieval Group Milliyet test collection. For obtaining the indexing terms use two stemmers: a. Simple F5 (first five) stemmer and b. another stemmer (Porter's algorithm, Zemberek, Successor Variety etc.). For the second stemmer you may use an implementation available on the web. You have to use a weighted approach for indexing, i.e., it cannot be binary indexing. Generate an inverted index structure separately for each stemmer.

   Note that each collection of size 100, 200, ... 1000 will have its own index, terms etc.

For indexing you may use a toolkit such as Apache Lucene, etc. For term weighting you may use the method suggested by the toolkit you use.

During indexing use the stop word list generated by our group given in Turkish IR paper.

2. Select two clustering algorithms of your choice. Explain why you select them. Preferably one of them is C^3M, note that C^3M is not mandatory. If you like you may introduce your own clustering algorithm and use it as one of the clustering algorithms. If you use an implementation available from the web make sure that you cite the source.

   Make sure that they are meaningful clustering algorithms none is a straw man. If you like you may modify the algorithm to resolve a weakness you observe.

3. Consider ten collections of sizes 100, 200, ..., 1000 increasing with increments of 100 documents. The first is the first 100 documents of the collection, the second is the first 200 documents of the collection and so on.

   Measure the average similarity among the members of the clusters (intra-similarity). Also measure the similarity among the member documents of different clusters (inter-similarity). For inter-similarity measurement you may use an approximate method based on cluster centroids.

4. Compare performance of the clustering algorithms using a statistical test to show which one is more effective, in other words generates more meaningful results. For meaningful clustering structures we expect to observe high intra-cluster similarity and low inter-cluster similarity. You may use them separately or use a F measure like approach and combine them.

5. Find an appropriate statistical test to compare the effectiveness of clustering algorithms with different indexing options.

   Note that this test involves comparison of two clustering algorithms (CA1, CA2) based on indexing structure with two stemmers (ST1, ST2). Therefore, it involves the comparison of four experimental results CA1-ST1, CA1-ST2, CA2-ST1, CA2-ST2 with 10 results for collection sizes of 100, 200, ... 1000.

6. Compare efficiency of clustering algorithms in terms of cluster generation time using a statistical method. Here consider only the execution times for clustering, without indexing time.

7. Write a paper to present your results. It must have the sections of a typical CS paper presented by using the ACM Conf. format. Make sure that it has an introduction, related work, methods used for clustering; effectiveness measurement, data set, computer system description, experimental evaluation based on statistical tests. The project is open to your imagination. Make sure that you have at least one extension other than what is defined here. Possibilities include the effect of no use of stopword list, effect of similarity measure used during clustering, effects of different indexing such as binary indexing, Monte Carlo experiments on effectiveness etc.

Bibliography (with brief notes)

1. Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M. "Information retrieval on Turkish texts." *Journal of the American Society for Information Science and Technology*. Vol. 59, No. 3 (February 2008), pp. 407-421.

   (Explains the construction of the largest Turkish information retrieval test collection that includes more than 400,000 documents and 72 documents with their relevant documents. Provides a stopword list of length 147 words.) From github try: BilkentInformationRetrievalGroup/IRCollection

2. Croft, W. B., Metzler, D., Strohman, T. *Search Engines: Information Retrieval in Practice*. 2005. (Free dowload: https://ciir.cs.umass.edu/irbook/. Explains different aspects of Galago.)

3. A tutorial of Galago: https://github.com/jiepujiang/cs646_tutorials (Java-based search engine based on Lemur and Indrie projects.)

4. Lemur Language Modeling Toolkit (4.12): http://www.lemurproject.org/lemur/ (Good for clustering)

5. Apache Nutch https://wiki.apache.org/nutch/FrontPage#Nutch_2.X_tutorial.28s.29 (Good for indexing)

6. Lucene http://lucene.apache.org/ (Open source toolkit includes the Indri C++ -based search engine, good for seeing document index vectors)

7. On cover coefficient clustering C^3M see: https://github.com/trinker/hclustext (Note: How complete I don't know.)

Groups

No.	**Members** **First person in each group is the team leader**
1	Akifhan Karakayalı Hamdi Alperen Çetin Yusuf Sait Canbaz
2	Ahmet Eren Başak Duygu Durmuş Muhammed Çavuşoğlu
3	İlayda Beyreli Mubashira Zaman Oğuzhan Karakahya
4	Furkan Akdemir Onur E. Karakaşlar
5	Alper Eroğlu Gizem Çaylak
6	Burak Mandıra Diyala Erekat
7	Berkay Taş Berksu Kısmet
8	Abdullah Kaan Karaata Mehmet Faruk Ongun Orhun Çağlayan
9	Undecided Students