Major Projects

Information Retrieval and Extraction [CSE474]

Major Projects

Twenty projects

Two teams per projects

Preferences for each group to be mentioned in the Google Form that will be shared soon

Begin communicating with your assigned mentors as soon as the projects are assigned

Major Projects

Deadline 1 : Project Scope Submission	March 9
Deadline 2 : Minimum Viable Product	March 26
Deadline 3 : Complete Deliverables	April 16

End Deliverables (Phase 3)

Complete End to End System (with code)

Project Report

Presentation Slides

3 Minute Video introducing your project's salient features and brief idea about the implementation

Further details will be shared later

Major Project Descriptions

#1: Document Summarizer

Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented,and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere"

Task: Given a document or a URL about a particular course get its summary (or snippet) in 2-5 lines.

Domain: Computer Science Courses

Challenges

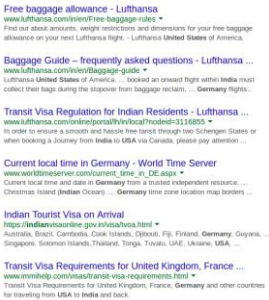
- Several existing approaches - which one performs the best.
- Does the approach vary w.r.t the subject under consideration.
- Parsing various file formats - Have to try various implementations available on the web. E.g. JSoup parser, Apache Tika, etc.
- Ensure that the summary is grammatically correct.
- Ensure that the summary is coherent and understandable.
- How to assess the quality of the summary
- Identify important aspects of a document

References

- Text Summarization Model based on Maximum Coverage Problem and its Variant
- Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression
- SIMFinder : A flexible tool for summarization
- G Flow :http://knowitall.cs.washington.edu/gflow/

#2: Set Expansion

India
USA
Germany



India
Pakistan
Afghanistan
USA
Germany
Russia
Australia
...

Domain: Computer Science Courses

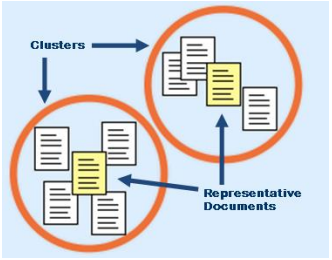
Challenges

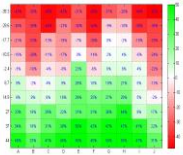
- How to extend traditional set expansion techniques to course related concepts.
- How do you filter out concepts related to a different course.
- Experiments need to be carefully designed because number of free API calls are limited.
- Precision vs Recall tradeoff.
- How to use other semantic similarity measures (if available) in addition to your approach
- When do you know that the set is full.

References

1. A Cross-Lingual Dictionary for English Wikipedia Concepts
2. Automatic Named Entity Set Expansion Using Semantic Rules and Wrappers for Unary Relations
3. Automatic Set Instance Extraction using the Web
4. Entity List Completion Using Set Expansion Techniques
5. Language-Independent Set Expansion of Named Entities using the Web
6. SEISA: Set Expansion by Iterative Similarity Aggregation
7. Web-Scale Distributional Similarity and Entity Set Expansion
8. Identifying Sets of Related words from World Wide web.

#3: Document Clustering, Feature Aggregation & Visualization





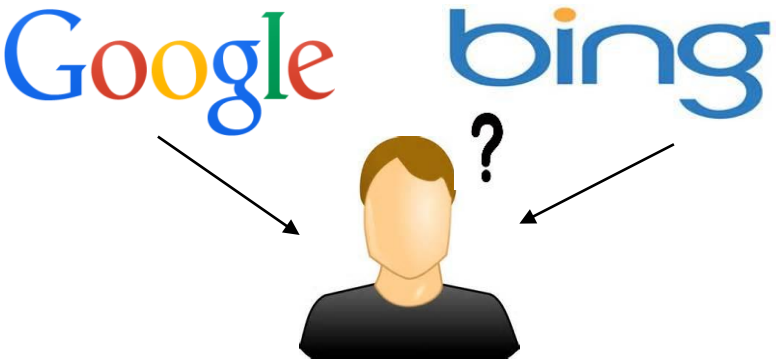
Challenges

- Parsing various file formats - Have to try various implementations available on the web. E.g. JSoup parser, Apache Tika, etc.
- Extract all possible features from text: NE, POS, TF, IDF, etc. - Apache UIMA in Eclipse
- Provide ability to cluster based on any set of selected features.
- Given 2 clusters: which are the most distinguishing features
- Need to explore various tools for generating graphs, heat maps, word clouds, etc

Prerequisites

- Ability to code in Java.
- Good Understanding of IRE course content taught till now.
- Interest in feature engineering and applying ML algorithms to solve various tasks.
- Interest in research problems related to IR / NLP.

#4: Search Evaluation and Visualization Framework



Task: Given two search streams or APIs, build a system for comparative evaluation of the search APIs.

Challenges

Involves significant engineering.
Requires understanding of search engines.
How cognitive bias may affect the judgements of human evaluators?
Best practices, code reusability and extensibility.

References

Search Engines: Information Retrieval in Practice: Bruce Croft
Likert Scale – en.wikipedia.org/wiki/Likert_scale
Chapter 4 from - http://arxiv.org/pdf/1302.2318.pdf

#5: KB querying Projects (KBs Freebase, YAGO, DBpedia)

Download and Index the knowledge base to query it graphically

Use case / Queries :

1. Does R exist in the KB
2. Does E exist in the KB
3. Given E and R, find all the triples
4. How many times is E in subject of triple?
5. How many times is E in object of triple?
6. Given a set of triplets, visualize it as a graph.
7. In the entity graph calculate min distance between entities (specify the resultant path)
8. In the entity graph calculate k nearest neighbours of an entity.

Challenges

1. Dataset is large
2. Data visualisation in terms of a graph is not easy

Pre-requisites

1. Java
2. Knowledge of sparql and rdf will help

#6: Named Entity Recognition in Twitter

Task: Named entity recognition is one of the first steps in most IE pipelines. The diverse and noisy style of user-generated social media text presents serious challenges, however. Performance still lags far behind that on formal text genres such as newswire. The goal of this shared evaluation is to promote research on NER in noisy text

Register, downlad the dataset and build system for the shared task “Named Entity Recognition in Twitter" at the WNUT http://noisy-text.github.io/index.html#

Participants teams are provided with training and dev data in addition to a baseline system.

Challenges

You are provided with dataset (Training and Testing) and a baseline system with baseline P, R and F.

So evaluation will be how much you beat the baseline.

2. Create / annotate a 500 tweet test dataset to prove your results.

References

Task description - http://noisytext.github.io/index.html#

Alan Ritter et al. Open Domain Event Extraction from Twitter KDD'12

K. Gimpel et al . Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments

#7: Identify salient named entity of a tweet. Evaluate with imaged tweet corpus

Problem : Identify the most important NE a tweet talks about.

This is referred as ‘whole tweet’ entity linking. This is of high demand in social media content analysis applications and tweet summarisation.

Challenges

1. The salient entity may or maynot be explicitly mentioned in the tweet.

2. Dataset : 5000 tweet corpus of tweets with images, annotated with salient entities. A row of this dataset will consist of the fields
<tweet><contained imageURL><NE1,NE2..><salientNE>
All four fields are mandatory
<tweet> contains tweet text
<image URL> URL of the image contained in the tweet
<NE> One or more named entities identified in the image
<salient NE> the NE the tweet is talking about

3. First identify the named entities a tweet talks about. Determine which among them is salient.
It might be possible that the salient entity is not present in the NEs you identified. Then how will you expand the your NEs set to capture the salient NE?

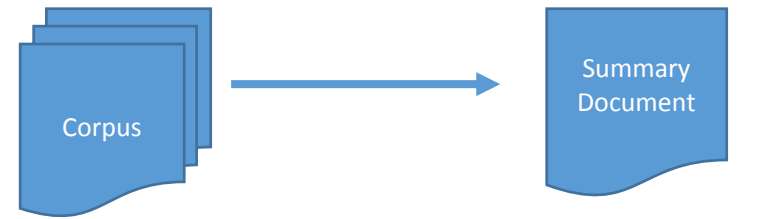
References

1. E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In
Proc. of the 5 th ACM Intl. Conf. on Web Search and Data Mining (WSDM) , pages 563–572.
ACM, 2012

2. Yamada et al. “Evaluating the helpfulness of linked entities to readers” HT'14

3. M. Gamon et al. “Identifying Salient Entities in Web Pages” CIKM'13

#8: Ordering LCUs to ensure best topical order



Corpus containing many documents

Single document with coherent sentences.

Local Coherrent Units

Ram lives in Hyderabad.
He loves Hyderabadi Biryani.
But he doesn't like Tandoori Chicken.

Sentences in a document acting as an LCU

The documents in the input corpus are segmented into a sequence of LCUs

Problem : To identify the best topical order for all LCUs in the corpora

Challenge : The LCUs may belong to different documents. The system should employ the best possible means to identify the topical order that is practiced in the corpus in each of the individual documents and find out the best possible global order for all LCUs in the corpus.

Problem Definition

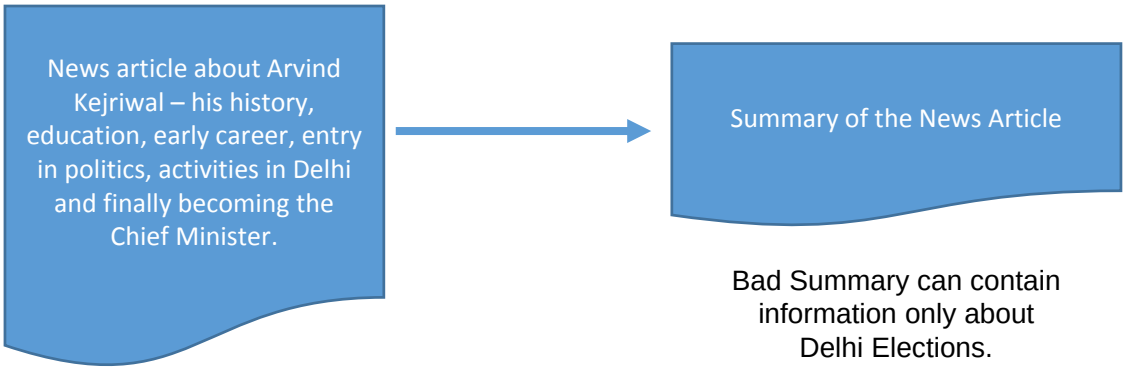
Input : set of topically related documents as a sequence of LCUs

Output : a topically order sequence all LCUs in the corpus

Note: Existing techniques of sentence ordering in multidocument summarization can be adapted and improved for this purpose

LCU*:”neighboring sentences grouped w r t the structural relations”

#9: Formulate a more reliable Topical Diversity measure for Text Summarization



Problem Definition

F(S) = Coverage (S) + Diversity (S)
length(S) <= k Bytes { $O(2^n)$ }

If F is monotonously non-decreasing submodular function, then there exists a greedy approach which creates a summary where $F(S_{greedy}) = k \times F(S_{optimum})$

A text summary is expected to exhibit more topical coverage with respect to source corpus and less redundant in terms of information conveyed.

Problem Definition

Problem:
Formulate a more reliable scheme to measure the diversity of summary which improves summary quality

Refer:
http://dl.acm.org/citation.cfm?id=2002537

#10: Community Detection

Key Terms : Graph Clustering, User Similarity, Centrality Measures

The problem is to work on user matrices to build weighted graphs upon which algorithms for detecting natural structure can be applied to identify compact communities.

Key Terms : User Similarity, Graph Clustering.

Details

Phase1 is to crawl a social network , and build user profiles and store interactions between them. We can use any social network or if a dataset is available use them directly.

Phase2 is to find communities in a social network based on:

1. Actions like commenting, likes, retweets etc.
2. The content they share or follow in a social network.

Phase3 will be to visualize the communities and come up with better evaluation techniques for the communities found in phase 2 .

#11: Computing social score of web artifacts

- Different social sites have different entities
  - Facebook - posts
  - Twitter - tweets
  - Youtube - videos

- Each of the sites have different metrics to calculate a social score (Its popularity),
eg: Facebook it would be likes

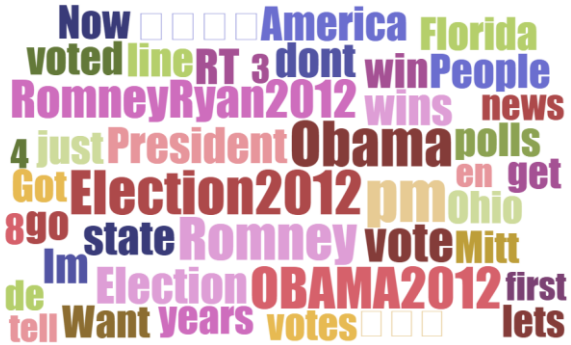
Details

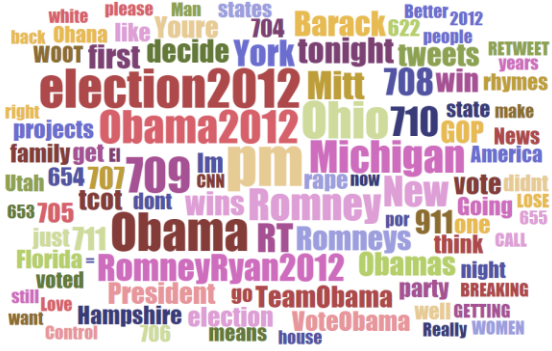
- Aggregate similar entities from multiple sources
- Associating a combined score obtained from each one of them.

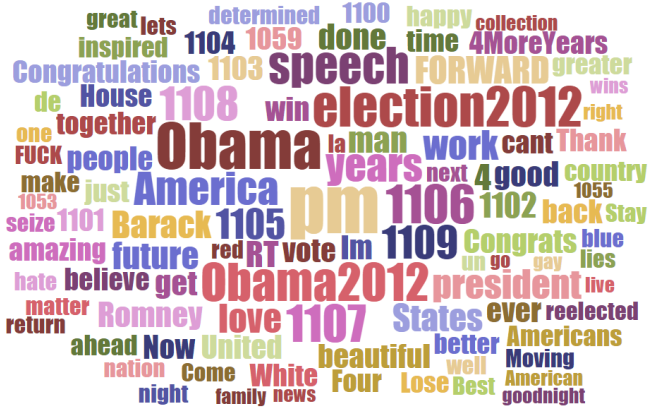
References
https://www.google.co.in/patents/US20070208583
http://pages.cs.wisc.edu/~tushar/projects/cs769.pdf for multiple sources

#12: Temporal Word Clouds

Problem Description







Challenges

Algorithmic

- On what basis are the words selected?
- How to extract temporality?
- How to score words?

Visualization

- How to present the word-cloud in pleasing manner so that it looks good as well as stays informative?
- How to place and orient words?

User Experience

- How to use animations to demonstrate the temporal nature of information?
- How to provide aesthetic interaction to the user to encourage exploration?

References

O. Kaser. Tag-cloud drawing: Algorithms for cloud visualization. In Proceedings of the World Wide Web Workshop on Tagging and Metadata for Social Information Organization

C. Wang. Importance-driven time-varying data visualization. IEEE Transactions on Visualization and Computer Graphics

Etiene Tiago. Linea: Tailoring timelines by visual exploration of temporal text

Weiwei Cui. Context preserving dynamic word cloud visualization

TagAssist: Automatic Tag Suggestion for Blog Posts. In ICWSM 2007

#13: Aspect Based Sentiment Analysis

The storyline was wonderful.

The plot was boring and I felt sorry for myself that I purchased the book.

The phone looks perfect but the camera is not flattering.

Problem Description

The **book** is an awesome read but they took 10 days to **deliver**.

Positive

Product

Negative

Delivery

References

Minqing Hu and Bing Liu. "Mining Opinion Features in Customer Reviews." Proceedings of Nineteeth National Conference on Artificial Intellgience (AAAI-2004), San Jose, USA, July 2004.

Minqing Hu and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, 2004.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05

Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)

http://alt.qcri.org/semeval2015/task12/

#14: Sentiment Analysis on Twitter

Input: Textual content of a tweet.

Output: Label signifying whether the tweet is positive, negative or neutral

Challenges

- Noisy text - Misspellings, lack of grammar

“@user9 wassup wid u??”

- Tweets are short - lack of context

“Saturday ill be at the LSU game with a ""fire Les Miles"" sign big faded”

- Open domain

“I had to wait for a long time” versus *“Laptop X has a long battery life”*

- *Sarcasm*

“Breaking News: Game of Thrones episode forces plane to make emergency landing #RedWedding”

Dataset

SemEval 2014 Task 10 subtask B provides both train and test sets of tweets along with labels here

#15: Predicting Potential Buyers and their buying behavior on ecommerce sites

Input:
A sequence of click events performed by some user during a typical session in an e-commerce website

Output:
Is the user going to buy items in this session?
If yes, what are the items that are going to be bought?

Challenges

1. No textual information is available. One needs to use the anonymised click through data in order to make predictions
2. Cold start problem
3. Sparse training data
4. Ongoing competition

Pre-requisites

1. Basics of SMAI
2. Interest in reading existing work

#16: Mining Opinions from Comparative Sentences

Input: A sentence that compares two entities
"Laptop X has longer battery life than Laptop Y"
Output: The entity preferred by the author
"Laptop X"

Challenges

- Relatively novel problem. Very little work done in the literature

Dataset

Download from here: Comparative Sentence Dataset

#17: Gender Detection in blogs

The goal of this project is, given a blog, you need to analyze the specific features in the text differentiating whether it is written by a male or a female.

The goal of this project is, given a blog, you need to analyze the specific features in the text differentiating whether it is written by a male or a female.

Details

The steps which need to be followed:

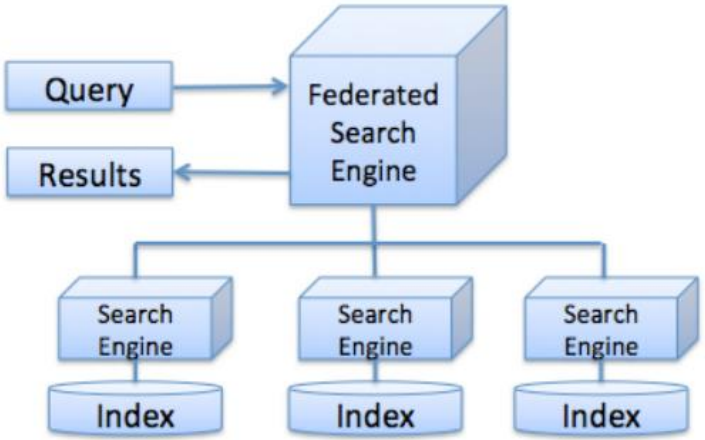
Dataset Collection → Features Extraction → Model Generation → Gender Identification

http://www.ccse.kfupm.edu.sa/~ahmadsm/coe589-121/cheng2011-gender-identification.pdf

#18: Federated search
On the basis of user interests

Federated Search - Simultaneous searches of multiple searchable resources.

Use Case - We have access to multiple research websites, but each of them has a search engine of its own. The idea is to create one which can search all of them.



Details

- In this project, federate search based on user history.
- If user is interested in Cricket, Bollywood and Politics
  - Crawl datasets for each of these topics
  - Calculate interests of user in each topic
  - Rank the results based on the user's interest.
- The amount of duplication in results.should be minimized.

#19: Finding Aspects for Opinion Mining

The **pizza** was wonderful but I don't know why **they** were so slow.

How to extract the right terms?

- List of words?
- All nouns?
- Grammatical relations?

Problem Definition

Sentiment Analyzer tools require reviews/tweets related to an aspect to extract Opinion about that aspect among users.

The Goal of the project would be able to detect relevant aspects/topics being discussed in large corpus of reviews/tweets. These are later given as input to Sentiment Analyzer tools.

#20: Phrase detection from text corpus

Given a text corpus, the aim is to extract the “phrases” also known as collocations.

- Collocations refer to sequences of words which occur together more times than we would expect co-occurrence due to chance.
- Various methods to get the phrases; NLP Techniques and Statistical Methods.
- Study various methods and implement a few of them that you find suitable.

Three Rings *for the* ***Elven-kings under the sky****,*

Seven for the ***Dwarf-lords*** *in their* ***halls of stone****,*

Nine for ***Mortal Men doomed to die****,*

One for the ***Dark Lord*** *on his* ***dark throne***

In the ***Land of Mordor*** *where the Shadows lie.*

One Ring *to rule them all,* ***One Ring*** *to find them,*

One Ring *to bring them all and in the darkness bind them*

In the ***Land of Mordor*** *where the* ***Shadows*** *lie.*

Three Rings ***Elven-kings under the sky***

Dwarf-lords ***halls of stone***

Mortal Men ***doomed to die***

Dark Lord ***dark throne***

Land of Mordor

One Ring

Uses

- Query Optimization

  An index which has N words, there are potentially N2 bigram phrases and N3 trigram phrases and so on

- Once Phrase has been detected it can be further used to find sentiments, Emotions of the given corpus.

Challenges

- Studying and deciding upon various NLP and Supervised/Unsupervised Techniques.
- Dataset can be large and diverse.

References

- Phrase Detection in the Wikipedia

[http://link.springer.com/chapter/10.1007/978-3-540-85902-4_10]

- R.C. Murphy. Phrase detection and the associative memory neural network. Neural Networks, 2003. Proceedings of the International Joint Conference, 4:2599– 2603, 2003.
- Focused Access to XML Documents 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007