Web Search Engines — Project Proposal

Himaja Rachakonda, Anirudhan J. Rajagopalan
N14633788, N18824115
hr970, ajr619

March 21, 2016

1 Title

The project is to be titled "Review-Rehashed" will summarize the reviews of a product with respect to a particular feature.

2 Team Members

1. Himaja Rachakonda, N14633788, hr970@nyu.edu
2. Anirudhan J. Rajagopalan, N18824115, ajr619@nyu.edu

3 Objective

The aim of this project is to extract sentences from online reviews which discuss about the features of a product and present it to the user. For example for a search term, say, "LG G3, Battery Life" will list the excerpts from all reviews that discuss about the battery life of LG G3 phone. Depending on the computational complexity and performance of the search engines, we will try to expand the search feature to multiple features and combination of features.

4 Sketch of Architecture

The search engine will be built by using the following different components.

Web interface: A web interface for the user to issue the search queries. This will be an interface with just two text boxes. One for the product name and another for the feature name.

Index store: The index store consists of data which can be retrieved from Lucene. The data will be populated and updated periodically by a crawler job that looks for updates to the data. The data will be stored in lucene after preprocessing.

Crawler job: Given a set of seed url, maximum depth of search and maximum number of pages to fetch, the crawler job will fetch all the urls that are reachable from the seedURL and dump the data in a folder.

Preprocessing module: This module takes care of preprocessing the data as required by the NLP module and dumps the processed data into Lucene.

NLP module: This module will be invoked during the query time. This is essentially a machine learning model that was pre trained using the data collected by the crawler. We use this to rank and get the list of all matching sentences for a given query.

5 List of web resources

We are planning to scrape data from

1. Amazon.com
2. Bestbuy.com

in very narrow categories of products (say electronics, or books).

6 Technologies used

We are planning to use the following resources for building the search engine.

Programming Language : Java, J2EE

Web server : Apache Tomcat

Index Store : Apache Lucene

NLP library : Apache Mahout