

**CZ4034 Information Retrieval**

Group members:

Yeong Jin Zhi

Heo Hui Lin

Eimoh Thwel Khaing

Platform used to crawl: Scrappy

Website to crawl: Lazada

**Question 1 - What has been done:**

* Write script to crawl specific data (eg: name, brand…)
* Write to json file

**Question 1 - Agenda:**

* Let the script run through the website to crawl data:

From the general category page -> go to first item then go to next without visiting back the same item -> go to next after visited every item on the page

* Continue until the end (the last page has different number of items) <how to do>
* Store everything in database

Question 1: Explain and provide the following:

1. How you crawled the corpus (e.g., source, keywords, API, library) and stored them (e.g., whether a record corresponds to a file or a line, meta information like publication date, author name, record ID)

2. What kind of information users might like to retrieve from your crawled corpus (i.e., applications), with example queries

* Warranty of earphone
* Brand of wireless mouse
* Discounted price of asus laptop
* Rating of earrings
* Original price of lamp
* Number of people rate for skater skirt
* Return policy of pillow
* Economy delivery price of luggage
* Standard delivery price of samsung tablet
* Express delivery price of sony camera

3. The numbers of records, words, and types (i.e., unique words) in the corpus

Question 2: Perform the following tasks:

• Build a simple Web interface for the search engine (e.g., Google)

• A simple UI for crawling and incremental indexing of new data would be a bonus (but not compulsory)

• Write five queries, get their results, and measure the speed of the querying

Question 3: Explore some innovations for enhancing the indexing and ranking.

Explain why they are important to solve specific problems, illustrated with examples. Possible innovations include (but are not limited to) the following:

• Interactive search (e.g., refine search results based on users’ relevance feedback)

• Improve search results by integrating machine learning or data mining techniques (e.g., classification or cluster techniques)

• Go beyond text-based search (e.g., implement image retrieval or multimedia retrieval

• Exploit geo-spatial data (i.e., map information) to refine query results/improve presentation/visualization

• Others (brainstorm with your group members!!)

Question 4: Perform the following tasks:

• Motivate the choice of your classification approach in relation with the state of the art

• Discuss whether you had to preprocess data and why

• Build an evaluation dataset by manually labeling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80%

• Provide evaluation metrics such as precision, recall, and F-measure and discuss results

• Discuss performance metrics, e.g., records classified per second, and scalability of the system

• A simple UI for visualizing classified data would be a bonus (but not compulsory)

Question 5: Explore some innovations for enhancing classification. Explain why they are important to solve specific problems, illustrated with examples. Possible innovations include (but are not limited to) the following:

• Ensemble classification (e.g., leverage on multiple classification approaches)

• Cognitive classification (e.g., use brain-inspired algorithms)

• Multi-faceted classification (e.g., take into account multiple aspects of data)

Video Link:

Link for file with crawled data:

Link for file with source codes:

**Using senticnet to come up with polarity of comments:**

1. Crawl all comments except empty ones
2. Stemming, Lemmatization?
3. Do polarity analysis on the comments then come up with a score by dividing the intensity, etc by the number of comments.

a) Rank by **Comments Polarity**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Product | Comments | Number of comments | Polarity Value | Intensity of Polarity | Score |
|  |  |  |  |  |  |

b) Rank by **Ratings**

|  |  |  |  |
| --- | --- | --- | --- |
| Product | Comments | Number of comments | Ratings |
|  |  |  |  |

c) Rank by **Combined Comments Polarity and Ratings**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Product | Comments | Number of comments | Ratings | Polarity Value | Intensity of Polarity | Score |
|  |  |  |  |  |  |  |

d) Rank by **Price**

|  |  |
| --- | --- |
| Product | Price |
|  |  |

e) OPTIONAL - Rank by **Price, Combined Comments Polarity and Ratings (special feature)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Product | Comments | Number of comments | Ratings | Polarity Value | Intensity of Polarity | Price | Score |
|  |  |  |  |  |  |  |  |