News Search Engine Group Project

I Requirements analysis & Summary design（李昶春、黄宸宇）

II Detailed design & Coding

Crawler (蒋宁欢) *DDL ：06-Feb*

1. Identify news website (recommended: 10+website, 1 million+ news in total)
2. Data should be saved in the format of XML containing Title, Full context, Date & Time, Source, URL, One related image, Topic, ID etc.

Indexing & Search Models (李哲宗、黄宸宇) *DDL ：10-Feb*

1.Construct Inverted Matrix Based on Title and Content(positional) referring to CW1
2. Index is saved efficiently? Stop words there or not? Stemming applied or not?
3.Construct vector space for each doc based on TFIDF
4.How to identify natural language
5.User query length; Free text or Boolean; Proximity/Phrase search?

Retrieval Models (黄铮) *DDL ：10-Feb*

1. Which one to select (TFIDF? BM25? OR TFIDF&BM25)
2. Popularity & Timeline should be involved in and other Impact?
3. New novel model optimized for you task?

Recommendation (黄宸宇) *DDL ：12-Feb*

1. Compute similarity of docs according to vector space.
2. Find most 5 similar news for each new.

Database (陈梓珩) *DDL ：13-Feb*

1. Data Storage and Call

Interface (李昶春) *DDL ：13-Feb*

1. Design an interface (Web Design + UX + HCI)
2. How results will be displayed (Classify and Exhibit the news according to topic)
3. Heading of document? Snippet?
4. Network Server related.

III Additional (Depend on workload) *13 Feb-17 Feb*

1.More than one language
2.Add dataset
3.multimedia
4.Classification of results
5.Query suggestion
6.Spell checker

7.Evaluation System

IV System testing (李昶春、黄宸宇) *14Feb-17 Feb*

V Report Synthesis (黄铮、蒋宁欢) *16Feb-18 Feb*

6-8 pages for project description (explain each component in you project and how it works what method/tool used to implement)

VI Report revision (陈梓珩、李哲宗、李昶春) *17Feb-19 Feb*

1-2 pages: each member of the group should write a paragraph/section on his/her contribution clearly in the report. Which role was taken, and what work was done.

VII Presentation (黄宸宇)

VIII Submission *DDL: 21-Feb 2020 17:00*