The University of Aizu

Intelligent Information Retrieval and Text Mining

Project 3

Lecturer: Vitaly Klyuev

ID: M5128110

Name: Fu, Yu-Hsiang

Date: 2008/11/13

Project 3: Creating a summary for the text file.

Goal: To give you feeling of the problems to implement a piece of software generating a simple summary.

Test Text: The test texts are chosen from the BBC.com in November 5 and 12, Obama wins historic US election [1] (as appendix 1) and Texting bug hits the Google phone [2] (as appendix 2).

Environment: The environment is HP Compaq CQ45-101TX NB. The hardware is performed on Intel Core 2 Duo P7350 2.0G CPU, 4GB RAM and Microsoft Windows Vista Basic Home operating system. The software is performed on Java 1.6.0 Update 5 with Gel IDE Tool.

Screenshot:

The screenshot of Project 3 application:

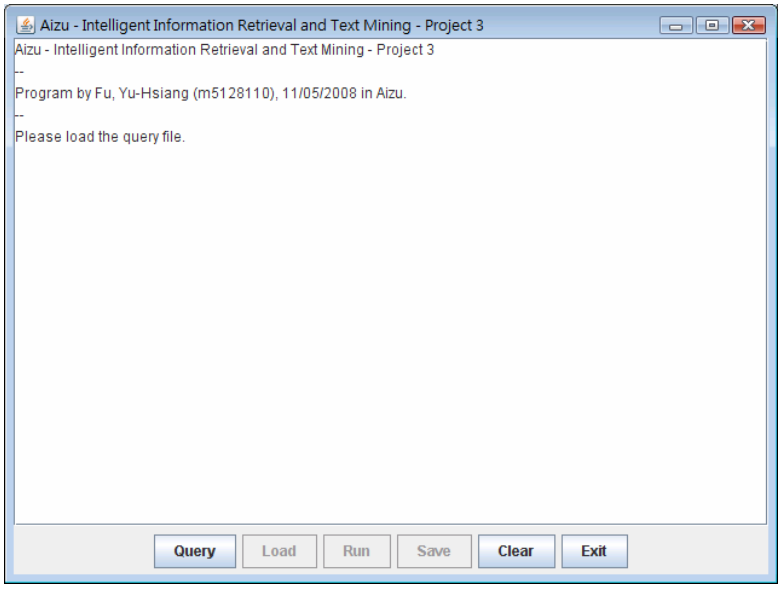


Figure 1. Screenshot of Project 3 application

Operation:

In this section, I'll try to describe all the operations of Project 3 application of Summarization briefly. Description will be explained by step-by-step of processes of the software. The Project 3 application can be executed by double-clicks on the Porject3.jar. The test text will be added the period in back of each subtitle of the news firstly. All the operations are as follows:

Step1. Click “Query” button to load the query file as figure 2(a).

Step2. Click “Load” button to load the text file as figure 2(b).

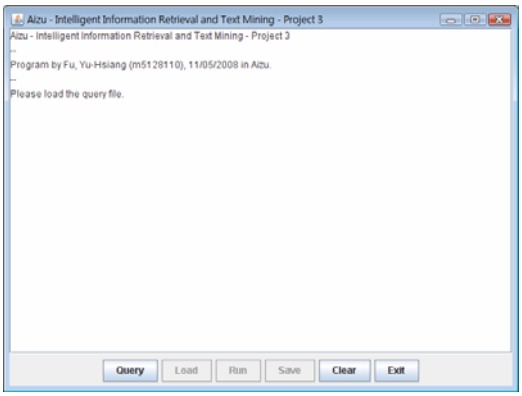
Step3. Click “Run” button to execute the program as figure 2(c).

Step4. The result will be shown in the display box as figure 2(d).

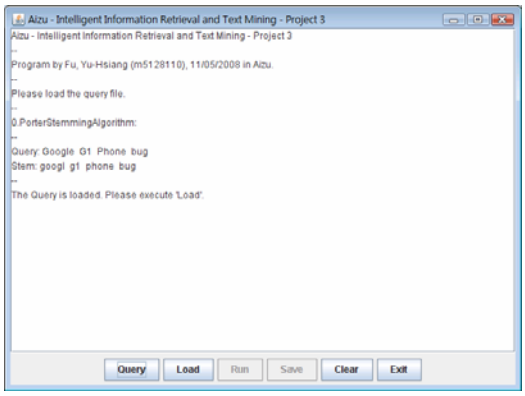
Step5. The result also can be saved using the “Save” button as figure 2(e). (If need)

Step6. Click “Clear” button to clean the display box. (If need)

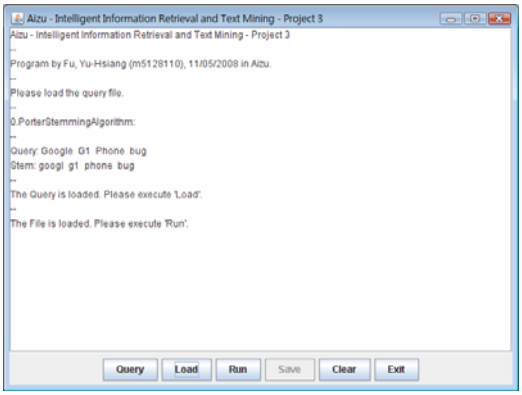
Result: The results both are 13 sentences of text 1 and text 2 (as appendix 3).



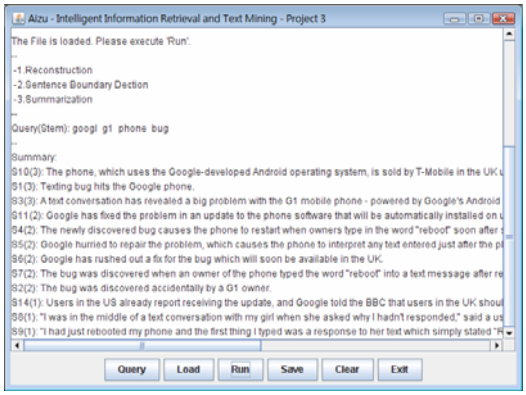
2(a) Load Query



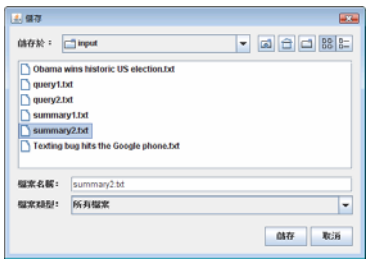
2(b) Load Text



2(c) Run



2(d) Result



2(e) Save

Figure 2. Operations of Project 3 application

Processes and Discussion:

In this section, it will talk about what the processes Project 3 application is. As figure 3, the first is input the query file and using the Porter Stemming Algorithm [3] to get the stem word of each query. The second is input the text, using reconstruction to reconstruct the text into the suitable form for next process of detecting the boundary between sentences. In the sentence boundary detection process, it will split the text into sentences which the boundary is detected. The summarization process will be executed which is matching the queries whether exiting in the each sentence is after the Porter Stemming Algorithm process and sentence boundary detection. Next, the results of the summarization will be ranked by the score of each sentence which is on the top with the highest score. The final is output the results of the summarization.

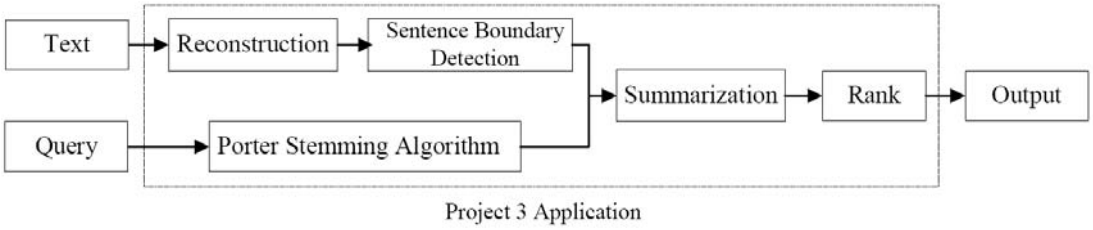


Figure 3. Processes of Project 3 application

The Porter Stemming Algorithm is proposed by Porter in 1980[3]; the stemming algorithm remove the suffix of the query words and keep maintaining the stem of words. In the Project 3, the stemming algorithm is used to get the stem word of the queries, and the summarization is used to match and retrieval the sentence which is matched by stem word of queries. The rank process is to calculate the score by a simple way which counts the number of time of matched keyword of each sentence.

For example, the queries are “Google”, “G1”, “Phone” and “bug”. The stems are “googl”, “g1”, “phone” and “bug” which are got by Porter Stemming Algorithms. The results of summarization are few sentences and already ranked are as follow:

1. The ***phone***, which uses the ***Googl***e-developed Android operating system, is sold by T-Mobile in the UK under the name "***G1***".
2. Texting ***bug*** hits the ***Googl***e ***phone***.
3. A text conversation has revealed a big problem with the ***G1*** mobile ***phone*** - powered by ***Googl***e's Android software.
4. …

The first three sentences contains three stem of queries, each sentence gets three points of score. The future work is that the Project 3 can be added the queries list which can let user to set up the weight or score of each query follow user’s requirement.

Reference:

[1] BBC News, “Obama wins historic US election,” http://news.bbc.co.uk/2/hi/americas/us_elections_2008/7709978.stm, 5 Nov, 2008.

[2] BBC News, “Texting bug hits the Google phone,” http://news.bbc.co.uk/2/hi/technology/7722367.stm , 12 Nov, 2008.

[3] M.F. Porter, ”An algorithm for suffix stripping,” *Program*, 14(3), pp 130-137, 1980.

Appendix 1:

The content of test text1:

Obama wins historic US election. Democratic Senator Barack Obama has been elected the first black president of the United States. "It's been a long time coming, but tonight... change has come to America," the president-elect told a jubilant crowd at a victory rally in Chicago. His rival John McCain accepted defeat, saying "I deeply admire and commend" Mr Obama. He called on his supporters to lend the next president their goodwill. The BBC's Justin Webb said the result would have a profound impact on the US. "On every level America will be changed by this result... [it] will never be the same," he said. Mr Obama appeared with his family, and his running mate Joe Biden, before a crowd of tens of thousands in Grant Park, Chicago. Many people in the vast crowd, which stretched back far into the Chicago night, wept as Mr Obama spoke. "If there is anyone out there who still doubts that America is a place where all things are possible, who still wonders if the dream of our founders is alive in our time, who still questions the power of our democracy, tonight is your answer," he said. He said he had received an "extraordinarily gracious" call from Mr McCain. He praised the former Vietnam prisoner of war as a "brave and selfless leader". "He has endured sacrifices for America that most of us cannot begin to imagine," the victor said. He had warm words for his family, announcing to his daughters: "Sasha and Malia, I love you both more than you can imagine, and you have earned the new puppy that's coming with us to the White House." But he added: "Even as we celebrate tonight, we know the challenges that tomorrow will bring are the greatest of our lifetime - two wars, a planet in peril, the worst financial crisis in a century. "The road ahead will be long. Our climb will be steep. But America - I have never been more hopeful than I am tonight that we will get there." Hours after Mr Obama's victory was announced, crowds were still celebrating in Chicago and on Pennsylvania Avenue in Washington DC. From red to blue. Mr Obama captured the key battleground states of Pennsylvania and Ohio, before breaking through the winning threshold of 270 electoral college votes at 0400 GMT, when projections showed he had also taken California and a slew of other states. Then came the news that he had also seized Florida, Virginia and Colorado - all of which voted Republican in 2004 - turning swathes of the map from red to blue. Several other key swing states are hanging in the balance. In Indiana and North Carolina, with most of the vote counted, there was less than 0.5% between the two candidates. However, the popular vote remains close. At 0600 GMT it stood at 51.3% for the Democratic Senator from Illinois, against 47.4% for Arizona Senator Mr McCain. The main developments include: Mr Obama is projected to have seized Ohio, New Mexico, Iowa, Virginia, Florida, Colorado and Nevada - all Republican wins in 2004. He is also projected to have won: Vermont, New Hampshire, Pennsylvania, Illinois, Delaware, Massachusetts, District of Columbia, Maryland, Connecticut, Maine, New Jersey, Michigan, Minnesota, Wisconsin, New York, Rhode Island, California, Hawaii, Washington, Oregon. Mr McCain is projected to have won: Alaska, Kentucky, South Carolina, Oklahoma, Tennessee, Arkansas, Alabama, Kansas, North Dakota, Wyoming, Georgia, Louisiana, West Virginia, Texas, Mississippi, Utah, Arizona, Idaho, South Dakota. Turnout was reported to be extremely high - in some places "unprecedented". The Democrats made gains in the Senate race, seizing seats from the Republicans in Virginia, North Carolina, New Hampshire, New Mexico and Colorado. They also increased their majority of the House of Representatives. Exit polls suggest the economy was the major deciding factor for six out of 10 voters. Nine out of 10 said the candidates' race was not important to their vote, the Associated Press reported. Almost as many said age did not matter. Several states reported very high turnout. It was predicted 130 million Americans, or more, would vote - more than for any election since 1960. Many people said they felt they had voted in a historic election - and for many African-Americans the moment was especially poignant. John Lewis, an activist in the civil rights era who was left

beaten on an Alabama bridge 40 years ago, told Atlanta's Ebenezer Baptist Church: "This is a great night. It is an unbelievable night. It is a night of thanksgiving." Besides winning the presidency, the Democrats tightened their grip on Congress. The entire US House of Representatives and a third of US Senate seats were up for grabs. Democrats won several Senate seats from the Republicans, but seemed unlikely to gain the nine extra they wanted to reach the 60-seat "super-majority" that could prevent Republicans blocking legislation.

Appendix 2:

The content of test text2:

Texting bug hits the Google phone. The bug was discovered accidentally by a G1 owner. A text conversation has revealed a big problem with the G1 mobile phone - powered by Google's Android software. The newly discovered bug causes the phone to restart when owners type in the word "reboot" soon after starting up the device. Google hurried to repair the problem, which causes the phone to interpret any text entered just after the phone was turned on as a command. Google has rushed out a fix for the bug which will soon be available in the UK. The bug was discovered when an owner of the phone typed the word "reboot" into a text message after restarting the phone. "I was in the middle of a text conversation with my girl when she asked why I hadn't responded," said a user called jdhorvat in the description of his discovery that was posted to Google's problem reporting website. "I had just rebooted my phone and the first thing I typed was a response to her text which simply stated "Reboot" - which, to my surprise, rebooted my phone." The phone, which uses the Google-developed Android operating system, is sold by T-Mobile in the UK under the name "G1". Google has fixed the problem in an update to the phone software that will be automatically installed on users' phones. "We've been notified of this issue and have developed a fix," said a Google spokesperson in a statement. "We're currently working with our partners to push the fix out." Users in the US already report receiving the update, and Google told the BBC that users in the UK should receive it by 12 November.

Appendix 3:

Results of test text 1:

1. Mr Obama captured the key battleground states of Pennsylvania and Ohio, before breaking through the winning threshold of 270 electoral college votes at 0400 GMT, when projections showed he had also taken California and a slew of other states.
2. Obama wins historic US election.
3. The main developments include: Mr Obama is projected to have seized Ohio, New Mexico, Iowa, Virginia, Florida, Colorado and Nevada - all Republican wins in 2004.
4. Democratic Senator Barack Obama has been elected the first black president of the United States.
5. Besides winning the presidency, the Democrats tightened their grip on Congress.
6. Mr Obama appeared with his family, and his running mate Joe Biden, before a crowd of tens of thousands in Grant Park, Chicago.
7. Many people in the vast crowd, which stretched back far into the Chicago night, wept as Mr Obama spoke.
8. Hours after Mr Obama's victory was announced, crowds were still celebrating in Chicago and on Pennsylvania Avenue in Washington DC.
9. "It's been a long time coming, but tonight... change has come to America," the president-elect told a jubilant crowd at a victory rally in Chicago.
10. Several other key swing states are hanging in the balance.
11. His rival John McCain accepted defeat, saying "I deeply admire and commend" Mr Obama.
12. It was predicted 130 million Americans, or more, would vote - more than for any election since 1960.
13. Many people said they felt they had voted in a historic election - and for many African-Americans the moment was especially poignant.

Results of test text 2:

1. The phone, which uses the Google-developed Android operating system, is sold by T-Mobile in the UK under the name "G1".
2. Texting bug hits the Google phone.
3. A text conversation has revealed a big problem with the G1 mobile phone - powered by Google's Android software.
4. Google has fixed the problem in an update to the phone software that will be automatically installed on users' phones.
5. The newly discovered bug causes the phone to restart when owners type in the word "reboot" soon after starting up the device.
6. Google hurried to repair the problem, which causes the phone to interpret any text entered just after the phone was turned on as a command.
7. Google has rushed out a fix for the bug which will soon be available in the UK.
8. The bug was discovered when an owner of the phone typed the word "reboot" into a text message after restarting the phone.
9. The bug was discovered accidentally by a G1 owner.
10. Users in the US already report receiving the update, and Google told the BBC that users in the UK should receive it by 12 November.
11. "I was in the middle of a text conversation with my girl when she asked why I hadn't responded," said a user called jdhorvat in the description of his discovery that was posted to Google's problem reporting website.
12. "I had just rebooted my phone and the first thing I typed was a response to her text which simply stated "Reboot" - which, to my surprise, rebooted my phone."
13. "We've been notified of this issue and have developed a fix," said a Google spokesperson in a statement.