Yonatan Kirby

[yonikirby@gmail.com](mailto:yonikirby@gmail.com)

http://yonatankirby.atwebpages.com/

Thank you for requesting my Wikipedia project. The project theoretically enables the user to ask a question of value, such as "How tall is the average giraffe?" or "How wide is a normal football field?" At this time, the project is functional with only some questions, but the code I have so far is fairly impressive in my opinion. Here is how the code works:

After asking the user to input a question, the program immediately determines the parts of speech for the words in the question, using the FastTag library. Let us assume the user inputs "How tall is the average giraffe?" Once the question is inputted, the program isolates three different words: the two adjectives in the sentence, "tall" and "average", as well as the noun, which is "giraffe". Then it queres Wikipedia for the article entitled "Giraffe". After this, it looks for "tall" and "giraffe" in the same sentence (although future versions should utilize a synonym-generator and search for words like "height" as well). After further narrowing down the candidate sentences by searching for numbers within them (future versions should include word representations, like "eighteen" for 18), the program compares the word "average" to adjectives in the two finalist sentences, which are: "Fully grown giraffes stand 4.3-5.7 m (14.1-18.7 ft) tall, with males taller than females." and "A newborn giraffe is 1.7-2 m (5.6-6.6 ft) tall.". In this case, "average" would be compared to both "grown" and "newborn". We compare them using a word-similarity metric, which computes the distance between two words as a decimal weight. In this case, the distance between "average" and "grown" as less than the distance between "average" and "newborn", so we get our answer: "Fully grown giraffes stand 4.3-5.7 m (14.1-18.7 ft) tall, with males taller than females."

Thank you for requesting my program. The libraries I use are not able to be used in a jar file, so I will show you how to set up my program in Eclipse.

1) Go here and download all files:

<https://drive.google.com/open?id=1ITDL1jMtOJfO4KS3CbRYP7yNoZ1oM7kK>

1) In a new workspace, select file->import->general->existing projects into workspace->click next->select archive file-> and import textproject.zip

2) Open the OurProject package.

3) Right click on the root of the project, "yoni1," and select properties (at the bottom of the menu). Select "Java Build Path" from the list on the left, and then click on the "Libraries" tab. Import the JawJaw and ws4j libraries by clicking "Add External Jars."

4) Go and copy lexicon.txt and paste it in the Eclipse file structure in the IDE into two different places: once in the same package as FastTag.java, on the same level-----and again, on the same level as pom.xml (at the bottom). Now run FastTag.java. The output should be many different parts of speech: namely,

"The/DT

ball/NN

rolled/VBD

down/RB

the/DT

street/NN

"

If, however, the output is all nouns, then we have failed to put the lexicon file in the right space.

5) Extract the txtWiki zip file. Then go to WikipediaProject.java. Go to lines 53 and 54. Adjust the paths to point to where your own query.js and txtwiki.js are (they are part of your extracted zip file of txtWiki).

6) Now run WikipediaProject.java as a Java Application. It asks you to enter a question. Enter in "How tall is the average giraffe?" and then press enter. It takes between five and ten seconds to fully compute. It should provide you at the end with this response: "Here's your answer: Fully grown giraffes stand 4.3-5.7 m (14.1-18.7 ft) tall, with males taller than females. " If this is not the answer, restart the program and it should work.

You can substitute the word "giraffe" for anything that doesn't have a Wikipedia disambiguation page. Run the program again. Enter the question: How tall is the average elephant? Your answer will be “Male African bush elephants are typically 23% taller than females, whereas male Asian elephants are only around 15% taller than females.” This is the wrong answer, obviously, and in the future I need to weed out answers that have numbers with a percent symbol next to them. But it is still interesting to see how the candidate sentences are whittled down. One of the other candidate sentences that the program chooses ends up being the right one.