L113: Coursework III (graded)

Simone Teufel

11/11/2011

1 Lexical Coherence/Discourse Segmentation

Many people are surprised about Hearst's (1997) finding that block comparison outperforms vocabulary introduction. Your task is to simulate both methods on a short example test from the BBC website (published yesterday, 10/11/2011), and verify which method works better on this test.

You can simulate the algorithm by program or by hand, whichever suits you better. You will not be punished for small errors introduced by either. For time efficiency reasons, I would however recommend that you perform at least the first steps automatically (up to point 4).

The only objective "right answer" available for the task is as follows: we assume topic boundaries coincide with headlines in the original text. However, as this is a short text, only one such topic boundary can be observed. Does any of the methods find the break in your simulation? Which method gets closer? (Alternatively, if you believe this gold standard is wrong, you are free to define your own.)

You should then write a report on your findings and observations, maximum 1 page, and attach your algorithms's intermediate results when you submit (whether they are printed, written or a mixture).

Detailed instructions:

1. Download file `asteroid.txt` from the L113 website.
2. Tokenise the text, i.e., separate words from each other and remove punctuation.
3. Find a stoplist on the web (any you like), and blank out stopped words.
4. Print or write all pseudosentences in such a form that you or your program can easily perform the two actions required at each window gap:
   - Identify new vocabulary (this can for instance be done with color pen)
   - Identify lexical overlap

5. When following the Hearst TextTiling algorithm, please use a block size of 2 pseudosentences, and a pseudosentence length of 20 tokens.

6. At each gap between pseudosentences, print or write your results – I would like to see which terms receive scores, not just the numerical scores. (Leave space in step 3 to do so, if you are working manually.)

7. Compile each method's overall results by comparing your results to the gold standard, which you can find in `gold.txt`.

8. Write a one page report: what worked, what didn't? Any ideas how these algorithms could be improved?

9. Submit the report together with your worked example of `asteroid.txt`

Note: The goal of this exercise is not the preprocessing, but the lexical cohesion issues. You are not expected to write a tokeniser or stemmer (let alone a perfect one), but to demonstrate some understanding of how the two methods are applied. Therefore, please seek a compromise between reasonable-quality processing and time-efficiency of processin. Whether or not to lemmatise/stem, for instance, is your decision.

asteroid.txt:

An asteroid that is 400m (1,300ft) wide has passed by Earth, much to the delight of astronomers. Although invisible to the naked eye, scientists said they spotted strange structures on its surface as it spun past at 30,000mph (48 280.32 km/h). Asteroid 2005 YU55's was the closest an asteroid has been to Earth in 200 years, according to Nasa. It is also the largest space rock fly-by Earth has seen since 1976; the next visit by a large asteroid will be 2028. The aircraft-carrier-sized asteroid was darkly coloured in visible wavelengths and nearly spherical, lazily spinning about once every 20 hours as it raced through our neighbourhood of the Solar System. Ron Dantowitz, the director of the Clay Centre Observatory in Massachusetts, followed the asteroid through a telescope. "We're tracking the asteroid itself, so the stars are moving by in the background and the asteroid is actually streaking by at about 30,000mph," he said. "As we track it, it looks like the stars are moving in the background and the asteroid is locked on in the centre view. "It's not so much that we can see it tumbling like a rock in space, we're examining it for the brightness and colour." Nasa said it had been no closer than 201,700 miles (324,600km), as measured from the centre of the Earth. The rock reached its closest point to Earth at 23:28 GMT on Tuesday. It will now trace a path across the whole sky through to Thursday. The asteroid often travels in the vicinity of Earth, Mars and Venus, but Nasa said this fly-by had been the closest the asteroid had come to Earth in at least 200 years. "This is the closest approach by an asteroid that large that we've ever known about in advance," said Lance Benner of Nasa's Jet Propulsion Laboratory. But he stressed that there had been no chance that the pass would be anything other than a close encounter. "2005 YU55 cannot hit Earth, at least over the interval that we can compute the motion reliably - which extends for several hundred years," he said. Instead, the pass gave astronomers a rare opportunity to study the asteroid in detail. In particular, two radio telescopes - the Goldstone Observatory in California, US and the Arecibo Observatory in Puerto Rico, US - tracked radio echoes off it in a bid to understand better what it is made of and how it is shaped. The precise details of the asteroid's path will also help scientists to predict where it will go much further into the future. Earth has several regular visitors like 2005 YU55 - most famously the Apophis asteroid. Apophis has in the past been claimed as a possible future impactor when it returns to our neighbourhood in 2029 and again in 2036. There is, according to the latest calculations, no danger from Apophis either. However, it will pass much closer to Earth on 13 April 2029 - at a distance of 18,300 miles (29,500km).