# ADSA Assignment 03 – News Analysis

As a team, you are working for a Korean online newspaper providing news in English for “expats”, i.e., foreigners living and working in Korea. Your boss had the following intuition: Korean news in English, e.g., koreatimes.co.kr, are lexically and semantically simpler than British or American news and this does not entice international expats, who rather prefer to read their own countries newspapers and remain ignorant about what is happening in Korea . Drawing more expat readers will yield more advertising revenues. According to your boss, to get more readers, Korean news in English should be lexically and semantically richer than what they currently are.

Before hiring new copy writers and journalists to create richer English content, however, your boss wants to have empirical evidence demonstrating whether they are right or wrong and to what extent. To do so, they asked you to provide a comparative study of lexical and semantic richness of news writing in British online news and Korean online news in English.

Your task is the following:

You need to compare the lexical richness/complexity of news articles in British online newspapers and Korean online newspapers to assess whether your boss is right or not.  
   
This requires 3 steps:

**Step 1:** get a sufficiently large amount of news text from online newspapers to get reliable results (getting text from online Web sites is usually called “scraping”).

**Step 2:** compute the value of some metrics about lexical richness/complexity in English and Korean news texts that you have scraped.

**Step 3:**  compare the results and give a (short, but informative) presentation to your boss.

Good luck!

More detailed instructions:

**Step 1:** your boss has provided you with a “bad\_html\_scraper()” to scrape text content from a news page. As you will notice, this scraper does the hard work that you do not understand (e.g., get html content from a url and parse it into a more manageable “string”). However, it scrapes not only the main content of an article, but also other unrelated content such as texts in sidebars, links etc. You should improve the output of the “bad” scraper to keep only the article content relevant for your analysis.

The bad scraper has been tested with articles from theguardian.com/uk and koreatimes.co.kr. In order to get reliable results, you should use at least 10 news articles from each Web site, possibly about different topics, for your analysis. You are free to analyse also other news Web site if you want to really impress your boss ☺

**Step 2:** You can start from the metrics that we have defined and developed during the lectures, but you can also design your own new lexical complexity metrics if you wish to impress your boss ☺

**Step 3:** Your demo should focus on (i) describing the analyse that you made (e.g., number of type of articles analysed, metrics used,…) and (ii) presenting the results to your boss.   
Then you will be also asked questions about the code that you have developed.

Instructions:

* You have to conduct this project in **groups of 2 or 3 people, groups must be the same as Assignment 2**.   
  You have to choose one member of the group to become “the captain”: the captain is the one who will have to make the submission on blackboard.
* You have to **submit the code** that you develop **on blackboard** by the **deadline of Sunday November 27th at 10pm**.   
  Please zip the folder “assignment03” on your computer and upload the zip file.  
  The submission area will be available in due time.
* You have to **give a demo** of your code to the TA, Ms Aikerim Orken. This has to take place in the period November 28 – December 2. A Doodle poll to schedule the demo will be provided in due time.   
  You are not allowed to change your code between the submission deadline and the demo (we will check!).   
  “Giving a demo” means to show the functionality of your software. So, you have to develop appropriate code in the main() of your application to showcase the implemented functionality. Failing to demonstrate the implementation of (some of the) functionality will lead to point deductions.  
  All group members must be present at the demo and must demonstrate that they know the code.