NLP session 07: Preparation for next session

- Read sections 1 and 2.3 of *Recent Advances in Document Summarization* (get preprint here)
- Download the DailyMail data from the CNN/DailyMail summarization dataset. I recommend downloading a (slightly pre-processed) version directly from here (second link): https://github.com/JafferWilson/Process-Data-of-CNN-DailyMail. Prepare the DailyMail data for processing, by writing a script that returns, for each story, (i) a tokenized version of the story itself and (ii) a tokenized concatenation of its highlights (end of each story)

Optional

There are two topics you could introduce next week. This would count toward your in-class participation grade. If you want to present one of them, announce this on this week's forum on Aula Global. In this way, others will know that the topic is already taken.

1. Give an introduction to TF-IDF (term frequency–inverse document frequency). Explain how it is calculated and the intuition behind its two components. Discuss a concrete numeric example. The Wikipedia page on this measure is a good starting point, as is An information-theoretic perspective of tf–idf measures by Aizawa. This counts toward two participation credits.
2. Read Extractive based Text Summarization Using K-Means and TF-IDF by Khan et al. and explain the intuition behind their k-means TF-IDF approach to summarization. This is a more complex topic, so it counts as your entire participation credit.