Answer real questions using Wikipedia content

Axel Didier, Frédéric Assmus, Pierre Goncalves







Summary

- Reminder of the process
- Current state
- Issues encountered
- Future work



Reminder on the process

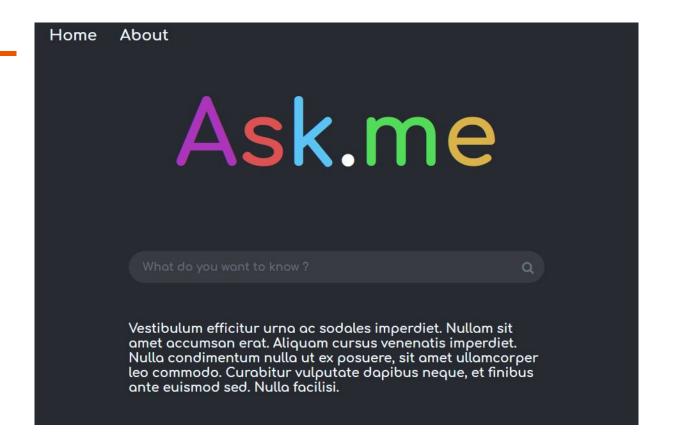
- **A.** Extracting the subject of the question
- B. Finding the most relevant wikipedia article given this subject+ the most relevant subpart of the article
- C. Giving the question and the subpart to BERT so it can find a pertinent answer
- The two first steps have to be finished before thinking about the third one



Current state

- **A.** Extracting subject with SpaCy to find nominal groups.
 - Keep other nominal groups as relevant information for the next step
- B. Wikipedia API functions to get the page associated with the subject
 - Currently, only keeping the summary of this article
- **C.** Cut the summary to fit BERT requirements
 - Check if the answer is coherent
- **D.** Display the answer: Jupyter widgets and a Webpage







Issues encountered for Subject Extraction

- Rule that we have for extracting the subject is really basic. It works for traditional questions like "When was the moon created?".
- Example of a question where we are not able to find the correct subject yet: "Who published The Lord of The Rings?".
- There is a lot of exceptions so we have to focus on finding new ways of extracting the subject



Issues encountered for Article Retrieval

- Issues at this step are directly related to our use of the API
- Sometimes we have weird results which compromise the process
 - o For somes cases we thought of using other tools like Request and Beautiful Soup in the future
- When disambiguation, the API didn't choose the most relevant article
 - We have to find a way to be sure that the chosen article will be relevant



Future work: priorities for the next 3 weeks

- Have a clean and concrete listing of the issues for the 2 first steps
- Try new solutions for subject extraction in order to avoid current issues
- Building a set of rules to be consistent in our wikipedia article retrieval
- Creating a wikipedia article splitting algorithm
- Go deeper in terms of rules and manage the exceptions better



Future work: if we complete the priorities

- Digging into papers to get ideas on how to make our BERT model work better and globally improve the other steps too
- Working on finding a way to evaluate our model
- Testing different ways of producing an answer? Instead of finding the perfect extract to give to BERT, why not giving multiple extracts?
- Extending to more than Wikipedia to match more types of question?



Thank you for your attention

Questions?