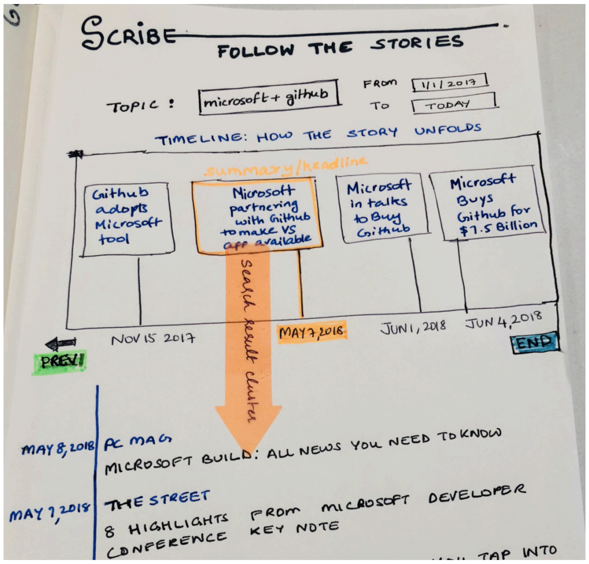
SCRIBE - news summarizer MVP

For this project, I am planning to build a news summarizer/ generate the most representative sentence from a news article, after selecting similar stories from different news stories in the dataset based on:

1. query combination
2. date


SCRIBE
FOLLOW THE STORIES
TOPIC : microsoft + github
FROM 1/1/2017
TO TODAY
TIMELINE: HOW THE STORY UNFOLDS
summary/headline
Github adopts Microsoft tool
Microsoft partnering with Github to make VS app available
Microsoft in talks to Buy Github
Microsoft Buys Github for $7.5 Billion
NOV 15 2017
MAY 7, 2018
JUN 1, 2018
JUN 4, 2018
PREV!
END
search result cluster
MAY 8, 2018 PC MAG
MICROSOFT BUILD: ALL NEWS YOU NEED TO KNOW
MAY 7, 2018 THE STREET
8 HIGHLIGHTS FROM MICROSOFT DEVELOPER CONFERENCE KEY NOTE
TAP INTO

Data

For the MVP, I have two large datasets that I am trying to examine. Both contain:

1. Date
2. url
3. headline
4. complete text
5. summary
6. author
7. source

Known unknowns:

1. LDA
2. choosing between summarizing and extracting most representative sentence - based on my specific use case
3. summarizing technique: extraction vs abstraction
4. have specific doubts on whether to use jupyter notebooks or set up GPU cluster on aws