Social media phenomena: Hate & Conspiracies
+ Ask DS question

Remaining Reddit data collection

IMT 547 - Social Media Data Mining and Analysis

26-Jan-2021 (Week 4, Day 7)

Today's Topics

- Project Idea generation
- Paper Discussion
- Asking DS questions (breakout rooms)
- Reddit data collection
- PS1 - out

Heilmeir's Questions

(As you start planning for your projects)

The Heilmeier Catechism

DARPA operates on the principle that generating big rewards requires taking big risks. But how does the Agency determine what risks are worth taking?

George H. Heilmeier, a former DARPA director (1975-1977), crafted a set of questions known as the "Heilmeier Catechism" to help Agency officials think through and evaluate proposed research programs.

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?
- What are the risks?

Heilmeir's Questions

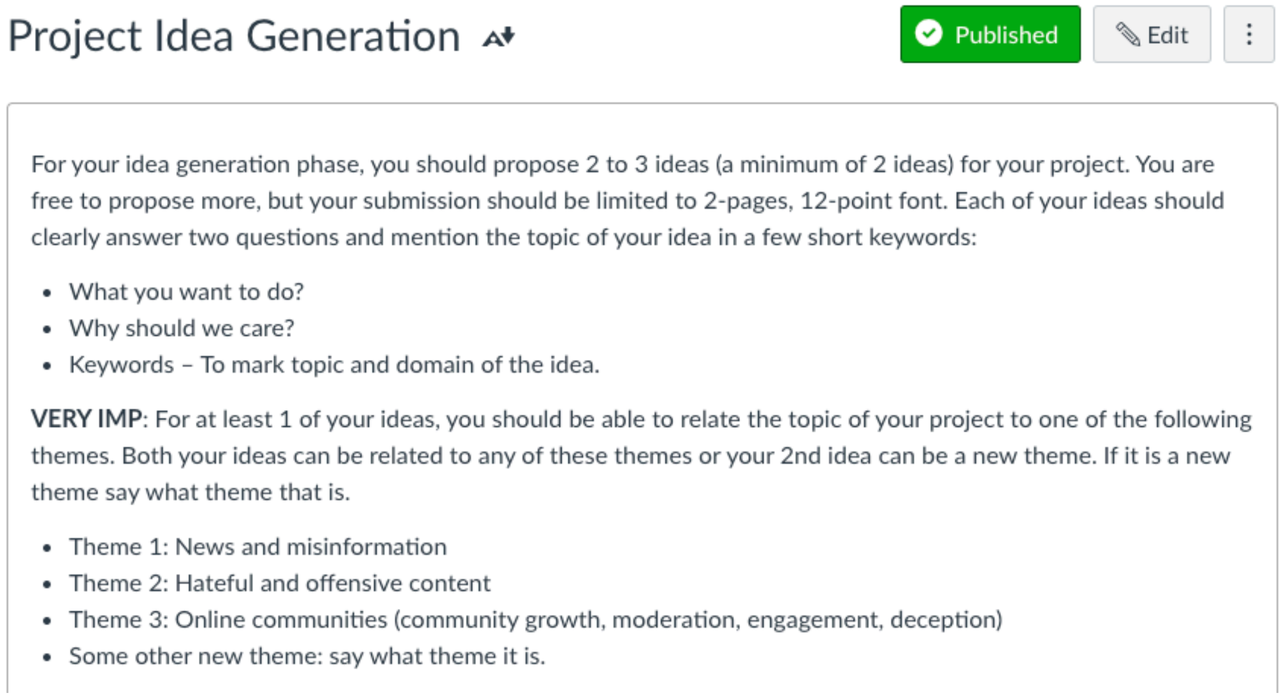
(Needed for project pitch)

1. **What are you trying to do?** Articulate your objectives using absolutely no jargon. What is the problem? **Why is it hard?**
2. **How is it done today**, and what are the limits of current practice?
3. **What's new in your approach and why do you think it will be successful?**
4. **Who cares?**
5. **If you're successful, what difference will it make? What impact will success have? How will it be measured?**
6. What are the risks and the payoffs?
7. How much will it cost (in terms of people, resources, etc.)?
8. How long will it take?
9. What are the midterm and final "exams" to check for success? How will progress be measured?

Project Idea Generation

Due this Friday (Grading: You will be graded for completeness.) This step helps in group formation!!

Check Canvas



Papers

Early Public Responses to the Zika-Virus on YouTube: Prevalence of and Differences Between Conspiracy Theory and Informational Videos

Adina Nerghes
KNAW Humanities Cluster
Amsterdam, The Netherlands
adina.nerghes@dh.huc.knaw.nl

Peter Kerkhof
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
p.kerkhof@vu.nl

Iina Hellsten
University of Amsterdam
Amsterdam, The Netherlands
i.r.hellsten@uva.nl

Automated Hate Speech Detection and the Problem of Offensive Language

Thomas Davidson,[1] **Dana Warmsley,**[2] **Michael Macy,**[1,3] **Ingmar Weber**[4]
[1]Department of Sociology, Cornell University, Ithaca, NY, USA
[2]Department of Applied Mathematics, Cornell University, Ithaca, NY, USA
[3]Department of Information Science, Cornell University, Ithaca, NY, USA
[4]Qatar Computing Research Institute, HBKU, Doha, Qatar
{trd54, dw457, mwmacy}@cornell.edu, iweber@hbku.edu.qa

Definition

Paper: What constitutes hate speech and when does it differ from offensive language?

both papers dealt with ambiguous topics that were difficult to both define and analyze.

- Jordan

Is it possible to even come up with a definition of hate speech that the majority of people would be satisfied with for regulatory purposes?

- Stephen

Borderline offensive language like using certain sensitive words is often debatable and that is what makes it sensitive

- Ayushi

Bias of human annotators

“Human coders appear to consider racists or homophobic terms to be hateful but consider words that are sexist and derogatory towards women to be only offensive, consistent prior findings” — Davidson Paper

the demographic of the human coders. Were the human coders predominantly men? Were the human coders randomly selected? What biases occurred because of the demographic of the human coders?

Joshua

Accuracy and reliability of CrowdFlower

Aniruddh

Policies & standards on hate speech

We also have a problem with different standards and policies on various online media. What is considered hate speech in one country might be different.

- Aftab

The complex nature of hate speech can be seen by how frequently websites like Twitter are updating their hate policy.

- Meghana

https://help.twitter.com/en/rules-and-policies#general-policies

General guidelines and policies

About **public-interest** exceptions on Twitter

COVID-19 misleading information policy

Violent threats policy

Abusive profile information

Glorification of violence policy

About **government and state-affiliated** media account labels on Twitter

About **search rules** and restrictions

About **rules and best practices** with account behaviors

Automation rules

About **Twitter's APIs**

About **Twitter limits**

Guidelines for Promotions on Twitter

Fair use policy

Counterfeit policy

Vine Camera Terms of Service and privacy policy

Updates to our **Terms of Service** and **Privacy Policy**

How to contact Twitter about a **deceased family member's account**

Defending and respecting the rights of people using our service

About specific instances when a **Tweet's reach may be limited**

Our range of **enforcement options**

Our approach to policy development and **enforcement philosophy**

Notices on Twitter and what they mean

Violent organizations policy

Twitter, our services, and corporate affiliates

Suicide and Self-harm Policy

Additional information about **data processing**

Abusive behavior

Automated copyright claims policy

Distribution of **hacked materials** policy

Platform manipulation and spam policy

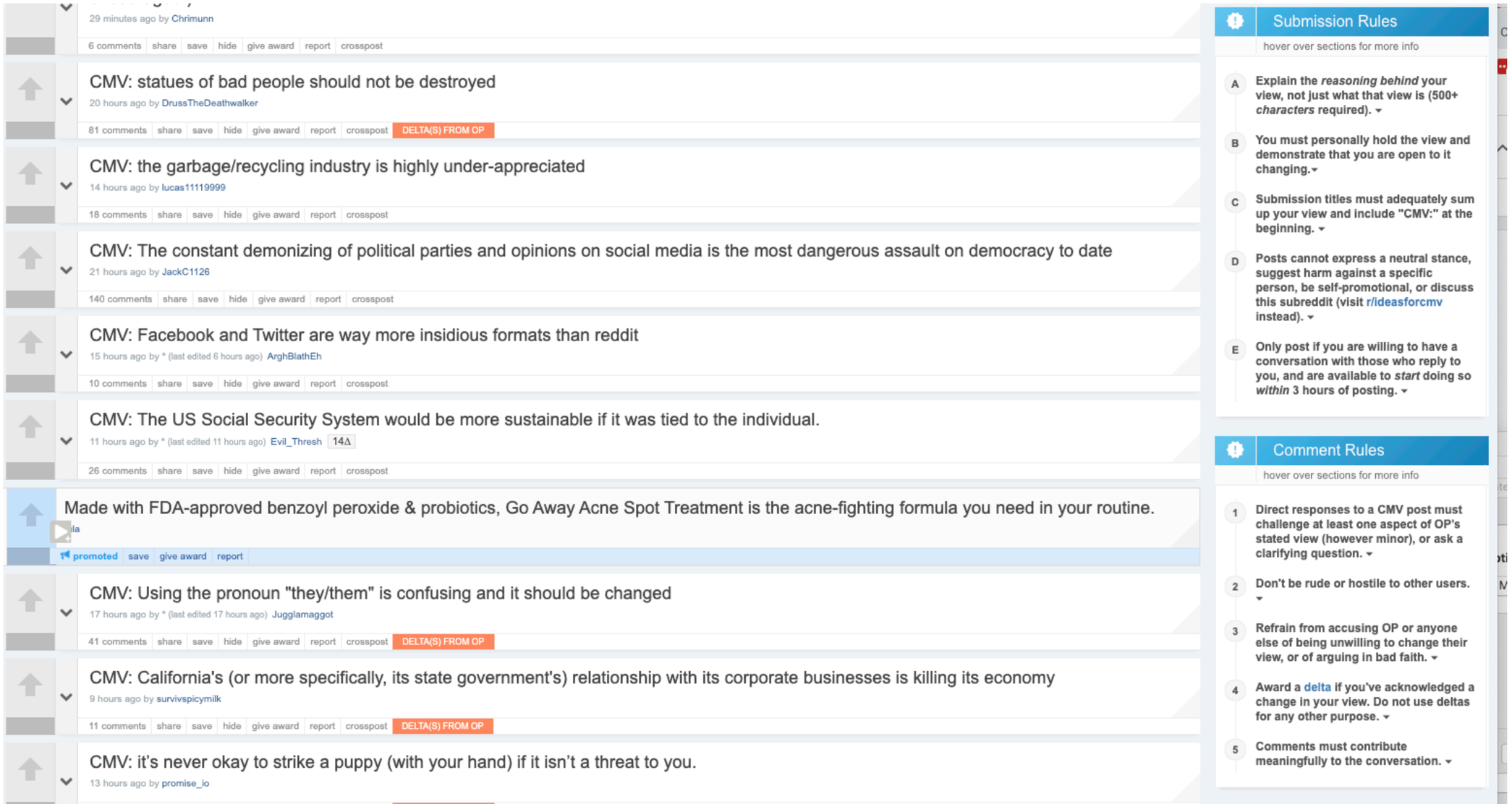
Reporting **false information** in France

Civic integrity policy

↑ Scroll to top

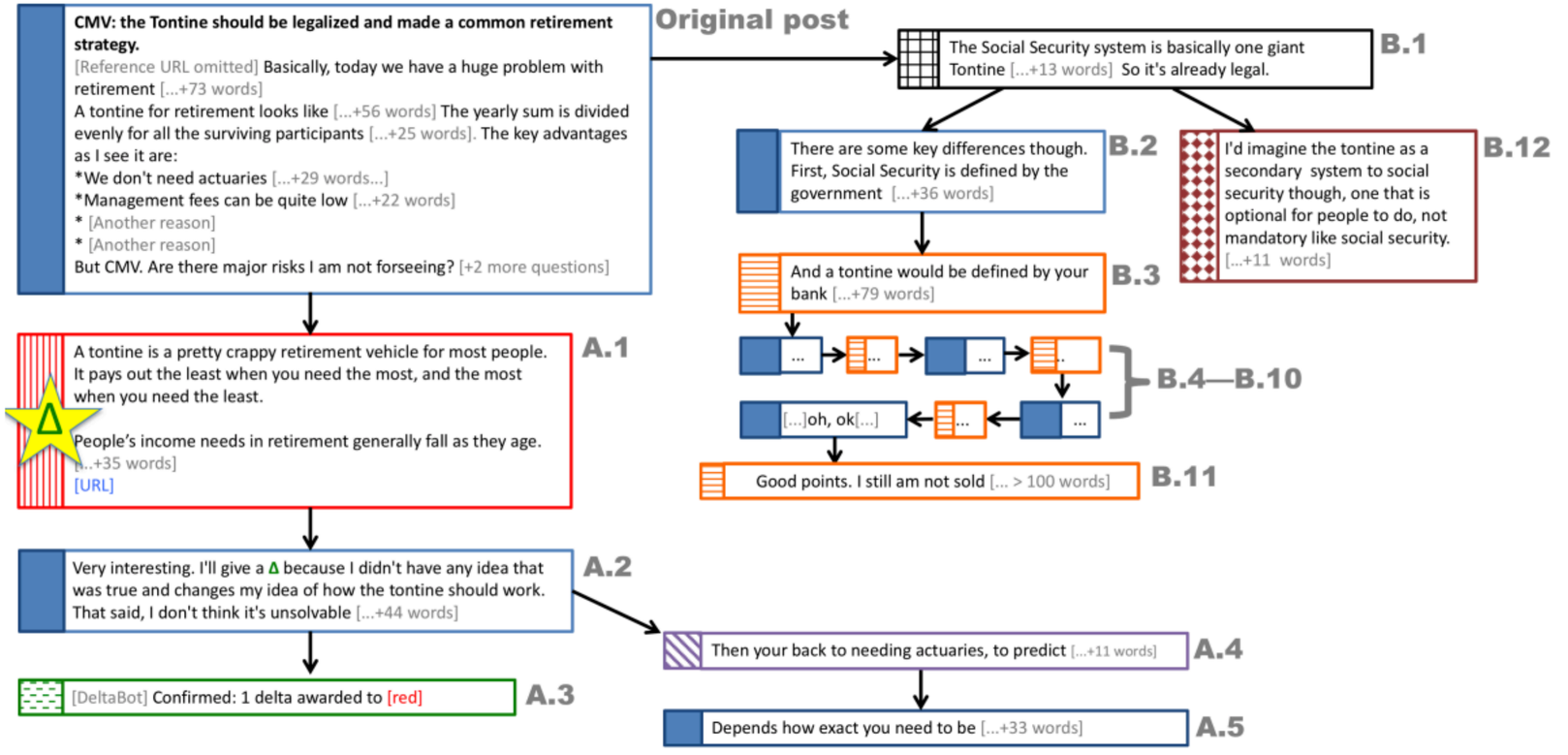
Community drive policies & rules

reddit.com/r/changemyview/



Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions

Chenhao Tan Vlad Niculae Cristian Danescu-Niculescu-Mizil Lillian Lee
Cornell University
{chenhao|vlad|cristian|llee}@cs.cornell.edu



DISCUSSION TREE

Figure 1: A fragment of a "typical" `/r/ChangeMyView` *discussion tree*—typical in the sense that the full discussion tree has an average number of replies (54), although we abbreviate or omit many of them for compactness and readability. Colors indicate distinct users. Of the 17 replies shown (in our terminology, every node except the original post is a reply), the OP explicitly acknowledged only one as having changed their view: the starred reply A.1. The explicit signal is the "Δ" character in reply A.2. (The full discussion tree is available at `https://www.reddit.com/r/changemyview/comments/3mzc6u/cmv_the_tontine_should_be_legalized_and_made_a/`.)

Method issues & suggested improvements

Many suggestions

some hate speech tweets are not in the form of text but some memes like static images and gifs. There might be some hate speech text on the images that cannot be collected by normal lexicon methods.

- Esther

t-davidson / **hate-speech-and-offensive-language**

Code | Issues 1 | Pull requests 1 | Actions | Projects | Security | Insights

master | 1 branch | 0 tags | Go to file | Code

master | hate-speech-and-offensive-language / **lexicons** /

t-davidson Initial commit

..

readme.md	Initial commit
refined_ngram_dict.csv	Initial commit

readme.md

Hate Speech Lexicons

This directory contains two lexicons that can be used to i

The file `hatebase_dict.csv` contains the original lexicon high recall it is associated with a high rate of false positiv manner (e.g. yellow, oreo, bird).

The file `refined_ngram_dict.csv` contains a refined lexi were contained in our labelled data and for each n-gram speech by the human coders. We then manually went thr

t-davidson Updated README `6d15050` on Jun 6, 2019 | **3** commits

classifier	Initial commit	2 years ago
data	Initial commit	2 years ago
lexicons	Initial commit	2 years ago
src	Initial commit	2 years ago
.gitignore	Initial commit	2 years ago
ICWSM_poster.pdf	Initial commit	2 years ago
LICENSE	Initial commit	2 years ago
README.md	Updated README	2 years ago

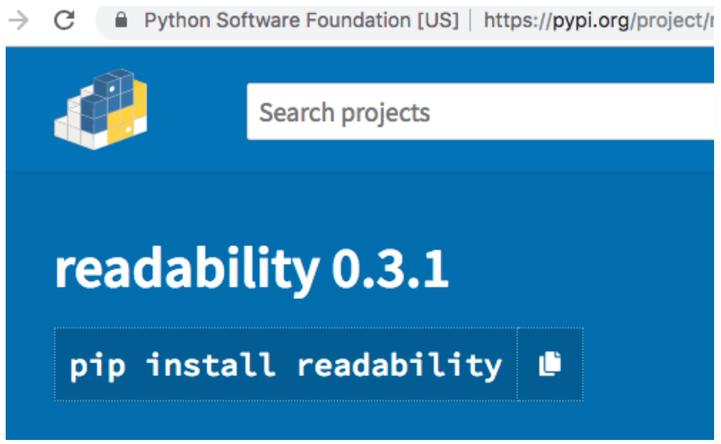
README.md

Automated Hate Speech Detection and the Problem of Offensive Language

Repository for Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM. You read the paper here.

Recall analytics techniques also used in previous papers

(will do notebook exercise today)




```
$ ucto -L en -n -s '' "CONRAD, Joseph - Lord Jim.txt" | readability
[...]
readability grades:
    Kincaid:                5.44
    ARI:                    6.39
    Coleman-Liau:           6.91
    FleschReadingEase:     85.17
    GunningFogIndex:        9.86
    LIX:                   31.98
    SMOGIndex:              9.39
    RIX:                    2.56
    DaleChallIndex:         8.02
sentence info:
    characters_per_word:    4.17
    syll_per_word:          1.24
    words per sentence:    16.35
```

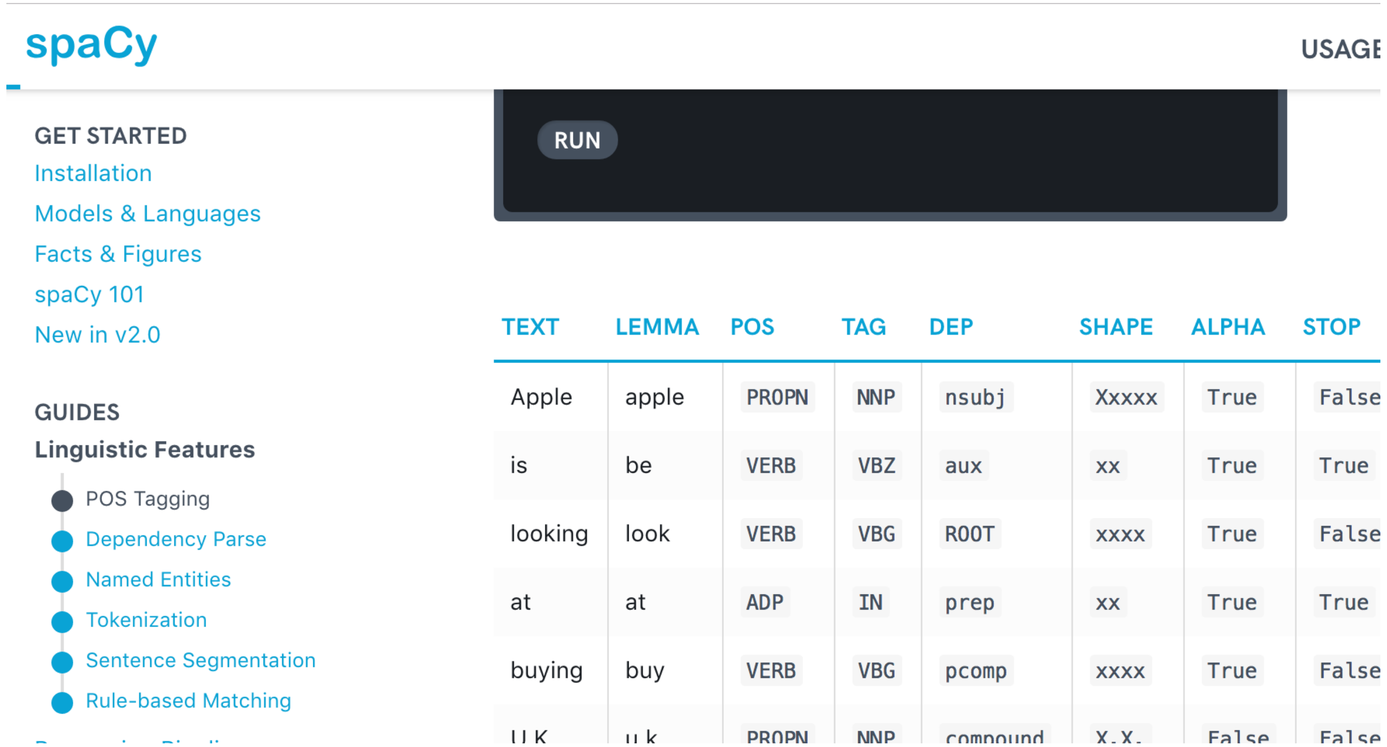

```
>>> textstat.flesch_reading_ease(test_data)
>>> textstat.smog_index(test_data)
>>> textstat.flesch_kincaid_grade(test_data)
>>> textstat.coleman_liau_index(test_data)
>>> textstat.automated_readability_index(test_data)
>>> textstat.dale_chall_readability_score(test_data)
>>> textstat.difficult_words(test_data)
>>> textstat.linsear_write_formula(test_data)
>>> textstat.gunning_fog(test_data)
>>> textstat.text_standard(test_data)
```

RUN

TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	VERB	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False

NLP techniques

Future labs on spacy and scikit-learn



Features:

POS, unigrams, bigrams…..

Asking DS questions!

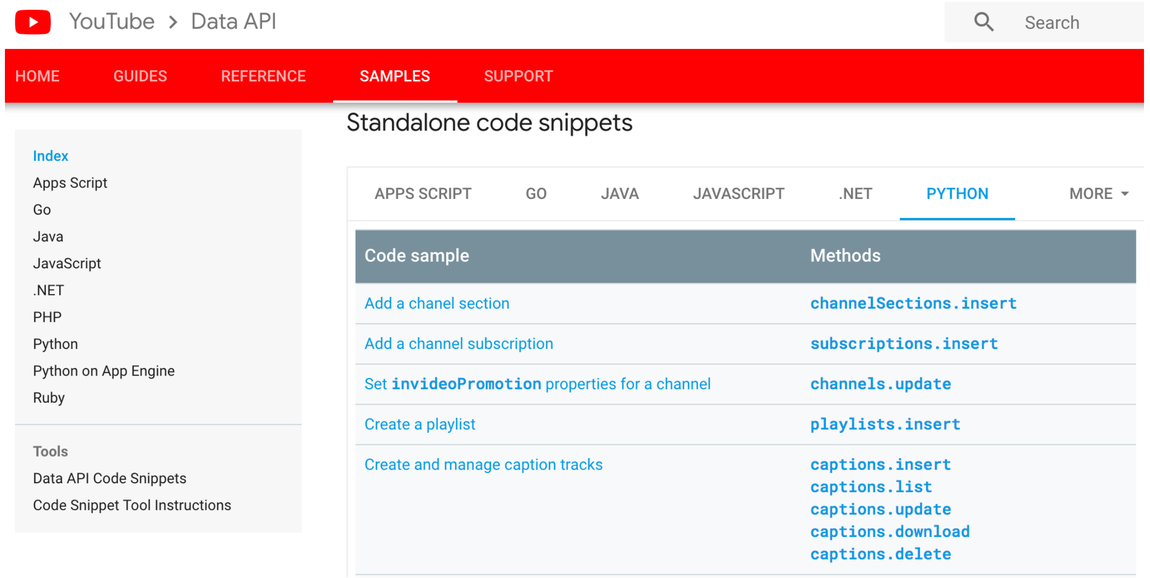
A). Data Sources: You have two data sources at your disposal:

1. YouTube Data API. Check all the ways you can get data from YouTube: https://developers.google.com/youtube/v3/code_samples/

2. hate speech lexicons from the paper that you read: https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/lexicons

B). Project Themes: Keep the following two project themes

- News and misinformation (*broadly defined*)
- Hateful and offensive content (*broadly defined*)



Q 1). Considering A) and B), what are some key Research Questions you can ask and how you can go about answering them?

Q2). If you had additional datasets at your disposal (Twitter, Reddit, news, Instagram, etc.), how will that help answer your RQs.

BREAK

See you at 9:45am

Leftover Reddit lab

Anonymous Feedback

https://forms.gle/Yp5suJ8c3qtGAArz9

Upcoming classes

Week 4 (Jan 25-29): Getting Started with Studying Social Media Phenomena

THU, JAN 28

Asking DS questions with social media

Exploratory Data Analysis

Required Reading

- Python Plotting for Exploratory Data Analysis

Lab: EDA

Idea generation writeup (due by Fri, 29-Jan)

Week 5 (Feb 1-5): Working with Data

Due: Problem Set I by 5pm, Mon Feb 1

As you think through project ideas, see lecture slides from previous classes for additional pointers