Creating word clouds with python

CodeUp X TechNomads
July 24th 2019

Kerry Parker

@_kaparker kaparker

Who I Am

Kerry - Data Scientist - PhD Physics

Analysis

Background in physics - involved data analysis, stats, visualisation

Code

Love learning new technologies and languages, primarily working in python

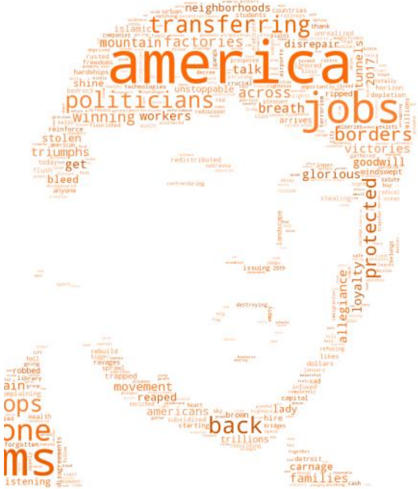
NLP

Recent work with NLP inspired personal project

Inspiration

- NLP project creating word clouds for visualising topic classification

- Came across inauguration word clouds



Word clouds

Visualisation of most frequent words in text

Application

- Vast amounts of open source datasets, eg. kaggle, github

- Initially planned ML predictor for GoT season 8 survivors…

- Incredible and distinct characters, interested in NLP for scripts

Getting Started

Following a similar method to inaugural word clouds, several steps involved when creating word clouds

1. **Finding relevant data**
2. **Cleaning data**
3. **Creating a mask from the image**
4. **Generating word clouds**

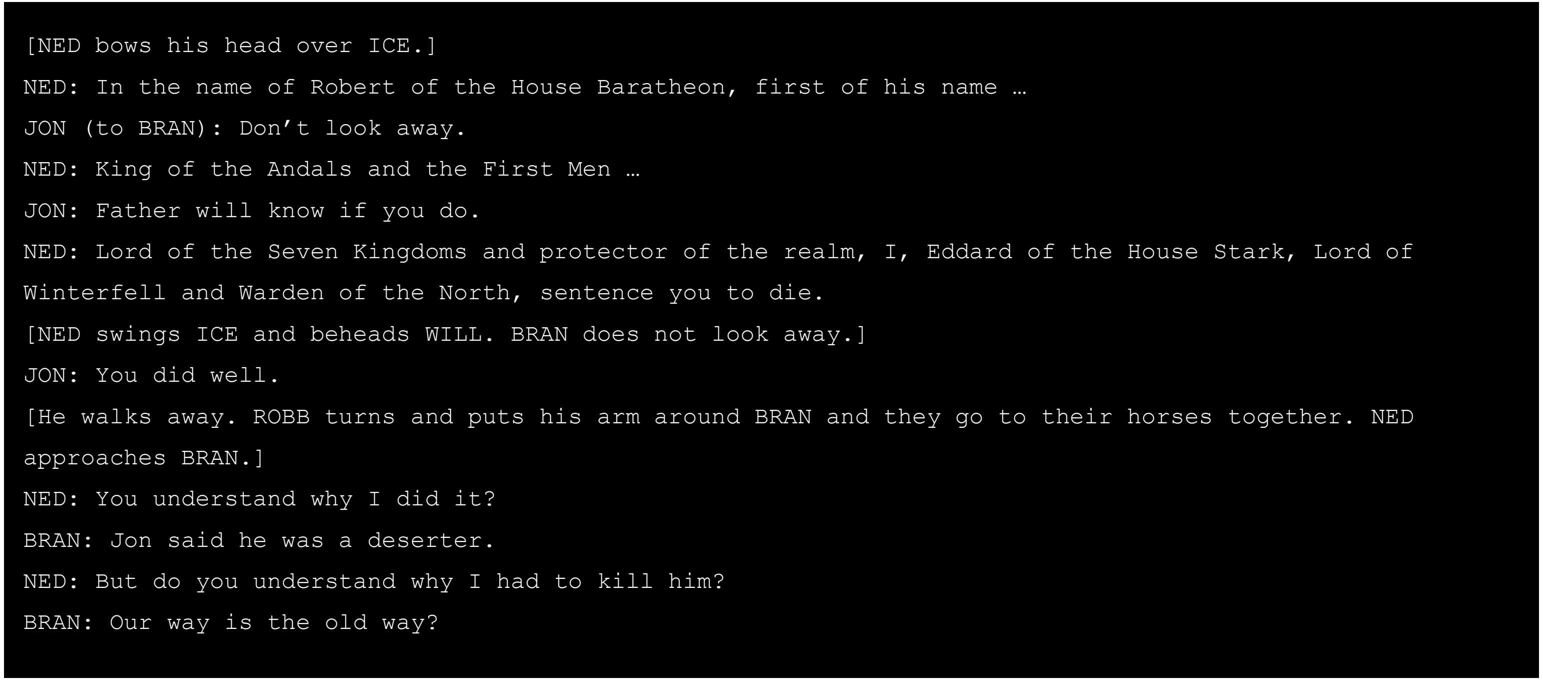
In this example I'm using python, can also use javascript, d3

Project goals: To create word clouds for GoT characters masked with image

1 - Finding relevant data

- Plenty of Game Of Thrones data available, scripts on GitHub
- Need each lines for each character… is this available?

1 - Finding relevant data



[NED bows his head over ICE.]

NED: In the name of Robert of the House Baratheon, first of his name …

JON (to BRAN): Don’t look away.

NED: King of the Andals and the First Men …

JON: Father will know if you do.

NED: Lord of the Seven Kingdoms and protector of the realm, I, Eddard of the House Stark, Lord of Winterfell and Warden of the North, sentence you to die.

[NED swings ICE and beheads WILL. BRAN does not look away.]

JON: You did well.

[He walks away. ROBB turns and puts his arm around BRAN and they go to their horses together. NED approaches BRAN.]

NED: You understand why I did it?

BRAN: Jon said he was a deserter.

NED: But do you understand why I had to kill him?

BRAN: Our way is the old way?

1 - Finding relevant data

- Plenty of Game Of Thrones data available, scripts on GitHub
- Need each lines for each character… is this available?

YES

...BUT, looking at several episode scripts the format varies

Uppercase or letter case, full name or first name, stage directions

1 - Finding relevant data

- Plenty of Game Of Thrones data available, scripts on GitHub
- Need each lines for each character… is this available?

YES

...BUT, looking at several episode scripts the format varies

Uppercase or letter case, full name or first name, stage directions

SOLUTION: user regular expressions with character names

Eg. `re.findall(r'(^'+char+r'.*:.*)', line, re.IGNORECASE)`

2 - Cleaning data

My approach:

- Remove line info eg. JON:
- Remove brackets - remove any stage directions from the character lines
- Remove accented characters and normalise
- Expand any contracted words eg. don’t → do not, using library for this
- (Optional): apply lemmatisation where a word is stemmed to a word in the dictionary, eg. is, are → be

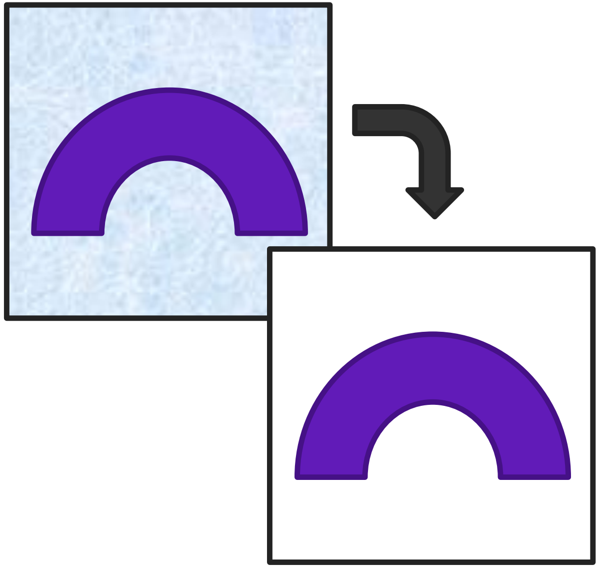
  Not currently using as doesn’t handle some words well eg. Stannis → Stanni
- Convert all text to lowercase
- Remove special characters (*,.!?)
- Remove stop words, using nltk stopwords

3 - Create a mask from image

- Based on inaugural word clouds, open images, create mask based on image colours or use grey function to present on black background

Points to note:

- Needed to remove background from image
- Image on white background preferable

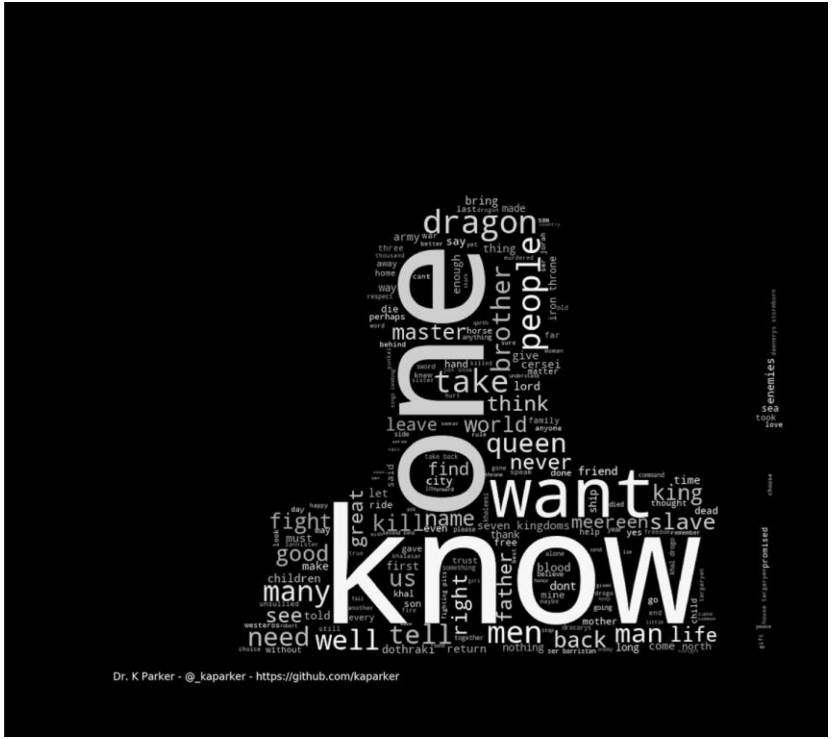


4 - Generate word clouds

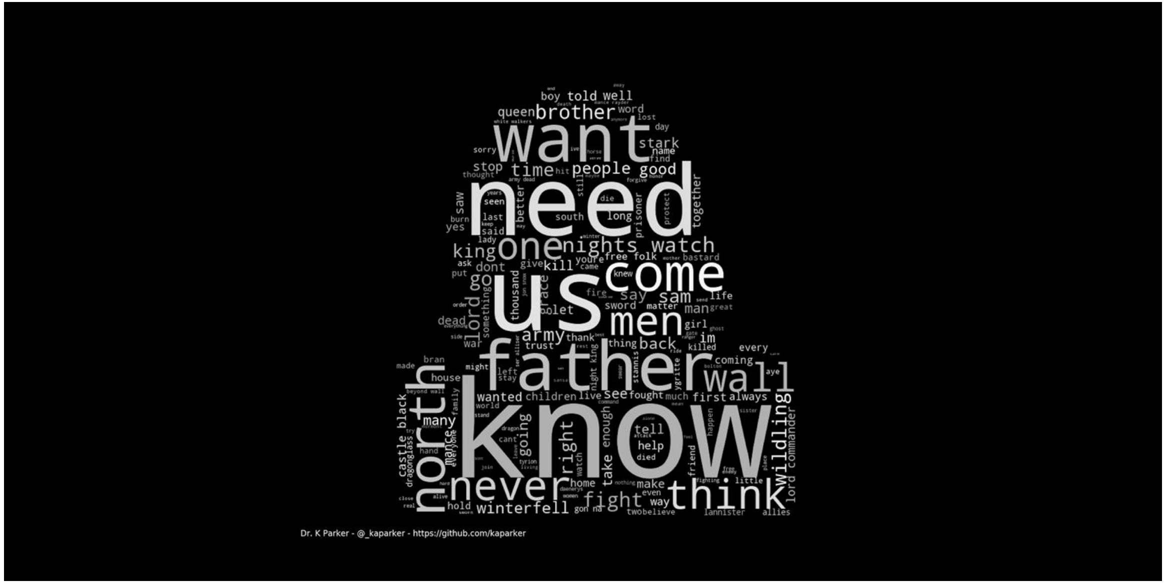
- Input words into word cloud, `generate()` function - output word cloud with size of text based on word frequency
- Can change `max_words, background_color` parameters to adjust
- Can `recolor` the image to either be grey scale or represent the image colours (or otherwise!)

Results

know
one
want
going
father
kill
lord
killed
see
say
girl
name
right
go
lady
good
brother
sword
got
dead
need
never
come
us
thing
friend
boy
joffrey
first
youre
mother
stark
little
take
Dr. K Parker - @_kaparker - https://github.com/kaparker


know
one
want
dragon
people
brother
take
queen
think
world
king
slave
meereen
man
life
men
back
tell
well
need
many
good
fight
kill
name
great
father
right
master
never
find
Dr. K Parker - @_kaparker - https://github.com/kaparker

Results

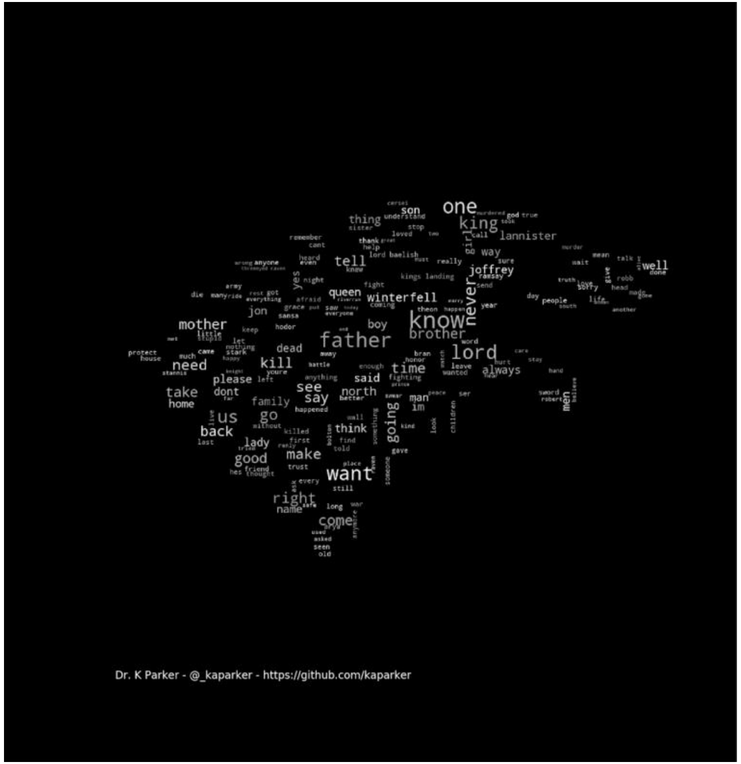


Dr. K Parker - @_kaparker - https://github.com/kaparker

Results

know
one
want
father
king
think
well
queen
good
brother
time
people
say
lannister
come
men
tell
us
yes
never
man
right
see
kill
love
need
son
sister
hand
make
thing
long
jaime
joffrey
family
lord
cersei
mother
little
take
war
back
going
believe
world
dragon
children
better
always
nothing
day
must
told
dead
much
boy
oh
go
let
course
find
head
killed
first
way

Dr. K Parker - @_kaparker - https://github.com/kaparker



Dr. K Parker - @_kaparker - https://github.com/kaparker

Improvements

Several ways can improve this!
Can use alternative techniques for text cleaning including:

- Different stemmer
- Add lemmatization
- Refine stop word list
- & more!

What’s next?

- Refine cleaning step – huge scope for improvement here!

- Further analyse text – get overall top words, who has the most lines per episode/season, who is the most mentioned character, generate new lines for characters…

- Find other films or shows to analyse - hunted for new scripts to analyse by character, currently no luck….

Questions?

dustin
yeah
oh
mean
friend
yeah yeah
okay okay
mike
time
really
good
day
find
go go
kid
mom
hawkins
hey
back
hell
one
see
make
jonathan
god
let us
thing
stop
guy
maybe
want
please
give
know
told
tell
okay
look
said
well
people
wrong
thank
kill
better
anything
got
nancy
found
even
hurt
come
night
thought
going
girl
talk
take
feel
uh
us
man
need
call
let go
understand
sorry
remember
way
right
gon na
steve
think
something
yes
home
say
little
still
sure
looking
nothing
listen
much

Links to projects

- Game Of Thrones Word clouds - https://github.com/kaparker/gameofthrones-wordclouds
- Stranger Things Word clouds - https://github.com/kaparker/stranger-things

- Get in touch on twitter or find me on LinkedIn if you have any questions or comments!