What I’ve had to do so far

Initially wanted to scrape PDF files from web links. Even worked on opening and reading such PDF files with fitz. But this was messy because certain scripts were in image form and not editable text form, so OCR was necessary. OCR could be done in python but unnecessary delay. Plus, all scripts from diff links, so not easy to automate the extraction of relevant text. Finally stuck to text files and switched back to read() inbuilt function because working with text file was easier for the following reasons:

* Could remove unnecessary preceding info to pilot script from txt file
* Can remove stuff from draft pilot which never made it to air (SS extra character, extra scene at end of pilot)
* Probably don’t have to wrangle with any random images etc. that might be present in the PDF form of the file

Save pdf to txt via avepdf, as adobe does not insert a space between words on preceding and subsequent lines that are part of the same sentence (so end-words get combined with 1st word of next line). Avepdf will also do OCR if the pdf doesn’t have recognizable text.

Individually loaded text from SS file, stored it as dataframe and saw if any issues so far. Also saw what basic cleaning could be needed.

Also done sent\_tokenize which will get individual sentences (it was able to render each sentence as 1st word character name, 2nd to last word character’s dialogue). Not sure what to do with this for now. Also, characters may speak in multiple sentences so this may not be sufficient. Will see after sentiment analysis.

Then automated this process so that all pilot text files in the folder can be processed in the same way (So far, SS & B99. Just have to update shows list for adding more shows)

Lot of issues with trying to store into list of text, \n\r chracters etc. which had to be addressed.

Finally after getting in dataframe form, calculated polarity & subjectivity of the 2 shows.

Plotted sentiment over time graphs (10-page chunks)

Now trying to calculate polarity & subjectivity of all **transcripts** (i.e. only of character names+dialogues) so that the setup/camera/etc.. info from the scripts does not contribute to the calculation.

Created document term matrix for word-related analysis. Reqd. additional cleaning (removing alphanumeric words and all numbers). Need to remove character names because that’s seeming most frequent. Or, maybe do an analysis of what each character says most? (word clouds for each character??)

Now focusing on trying to identify the major characters in the script using spaCy Named Entity Recognition (NER). Initially hoped to extract their dialogues.

Spent some trials on cleaning etc. Realized that removing punctuations has a negative impact on spaCy NER. Also had to do some script-specific interventions ($ signs in superstore script was causing weird italicization so replaced them with the word “dollar”; J.D. name in Scrubs script).

Used NLTK to create frequency distribution so that those entities that occur >2 are considered names of characters, because in each show certain names are part of the dialogue (usually pop-culture references/celebs) so those should not be considered characters. This was part of a function that identified entities of type ‘PERSON’. Freq distrib was organized in desc order so that I could spot char names fast instead of wading through many other words first.

Function was used because individual freq distribution was needed for each show.

Had to create a set of character names to avoid multiple entries for the same character.

Struggled with upper and lower case. Initially worked purely w lowercase script, but many names weren’t getting recognized in this form. Instead, opted for case-insensitive comparison with elements in the character-name set. Created a function for this because it was messy to execute case insensitive within token identification loop. Was also looking into truecaser to explore if adjusting casing to standard form would improve name identification, but realized that perfect name is of no use during dialogue identification, which is why I wanted the names.

Performed NER for one script then created function and loops to ensure that each script can be analyzed.

Initial approach for name recog:

for sent in doc.sents:

for tok in sent:

#look for names that have not already been identified. Otherwise each name gets appended to character-name list.

if tok.ent\_type\_=='PERSON' and tok.text in count\_dict.keys() and tok.text not in seen:

#print(tok.text)

char\_list.append(tok)

seen.add(tok.text)

return char\_list

Updated approach for name recog such that multi-word tokens can be considered (so that NER may have a better shot at identifying names)

for sent in doc.sents:

doc1=nlp(str(sent))

for tok in doc1.ents:

#print (tok)

#look for names that have not already been identified. Otherwise each name gets appended to character-name list.

if tok.label\_=='PERSON'

Was initially working on identifying full character names even if they’re multi-word names (e.g. Tom Haverford, Michael Scott). But realized (very late xD) that I don’t need that if I want to analyze character dialogues. Can straightaway look for sentences that begin with each of these entities, because that usually means that text between this and till the next name is the dialogue of that character. There will be some kinks but it may be a good start.

* Side note: Alarmingly the possessive form was being treated a separate name. Now that has stopped and I’m not sure how. Because when I tried “Amy Santiago’s”, it treated “Amy Santiago’s” as the word.

The list of char names is not organized in any particular order.

Intermediate approach list for which I don’t know what the code was so gotta redo and deal with bunch of crap words like a duffer:

30 rock

[LISA, Kenneth, PETE, FRANK, JOSH, SARIE, Gary, Jenna, JACK, LAWRENCE, Donaghy, TOOFER]

---

B99

[JAKE, Santiago, AMY, Morgenthau, CHARLES, Rosa, Rihanna, Peralta, Holt, Boyle, GINA, Clay, Diaz]

---

Its always sunny

[DENNIS, CHARLIE, TERRELL, Dee, Mac, KAREN, Janelle, BRETT, Paddys]

---

Mindy project

[Mindy, Castellano, Danny, Tom, Gwen, Dennis, Carl]

---

Parks and rec

[LESLIE, Ron, Tom, Barry, ANN, Mayor, Parks, ANDY, Aubrey]

---

The office

[Jim, Michael, Pam, Dwight, Oscar, Scott, Ryan, Roy]

---

Scrubs

[JAMES, KELSO, ELLIOT, TURK, JEFFERY, Bambi, Carla, Dorian, Reid, Pratt, Burski, BILLY, ROBERTS]

---

Silicon valley

[Thomas, DINESH, Steve, Boodle, Gregory, MONICA, ROLAND, LANGDON, Gavin, Jared, Belson, Peter]

---

Superstore

[Amy, Glenn, Dina, Jonah, Cheyenne, Frank, Mateo, Garrett]

---

Veep

[SELINA, Gary, Hallowes, Veep, DORSEY, Meyer, Dan, Anna, Frank, Reeves, Mike, Jonah, Tom]

---

Minimum revised version list:

30 rock

[LISA, Jenna, DeCarlo, KENNETH, Lemon, PETE, FRANK, Yeah, Sarie, Josh, Okay, Gary, Luderman, JACK, Donaghy, Lawrence, Jordan, Well, And, Hey, Yes, You, Oh, Hi]

---

B99

[JAKE, Detective, Santiago, TERRY, CHARLES, Okay, Captain, Holt, Gina, Rosa, Rihanna, Yeah, Great, Peralta, Meep, Yep, Sergeant, Um, Diaz, Boyle, Jamon, Iberico, Kane, AMY, Hello, Clay, Morgenthau]

---

Its always sunny

[DENNIS, Mac, CHARLIE, Sweet, Dee, TERRELL, Yeah, Oh, African, American, Yo, Dude, KAREN, Hey, Well, Janelle, BRETT, Look, Thanks, Alright, ?]

---

Mindy project

[Mindy, Tom, Castellano, Danny, Sorry, months, Gwen, Dennis, Carl, Springsteen, Okay]

---

Parks and rec

[LESLIE, Knope, Ron, Tom, Haverford, Barry, Laura, Linney, Ann, Mayor, Parks, Mark, Brendanawicz, Aubrey, SWANSON, Pit]

---

The office

[Michael, Jim, Oh, Yes, Scott, I, Pam, Dwight, Whassup, OK, Corporate, Jan, Um, Yeah, No, Todd, Packer, Hey, Uh, Stanley, Well, Oscar, Ryan, Howard, Watch, Dunder, Mifflin, Come, Assistant, Not, You, Roy, Why, Notes, Good]

---

Scrubs

[JAMES, Look, Turk, See, Kelso, Can, ELLIOT, Yeah, Good, Anyway, Your, Hey, JEFFERY, Reid, Dorian, CARLA, Bambi, Well, COX, Cmon, Dr., Uhhuh, You, Pratt, Tell, Burski, Yes, Okay, Great, Nice, BILLY, God, Ill, NURSE, ROBERTS, Oh, ,]

---

Silicon valley

[THOMAS, ERLICH, Steve, Jobs, GoolyBib, Fuck, Peter, Gregory, Pied, Piper, MONICA, ROLAND, LANGDON, Yeah, Peckering, Gavin, Belson, Hello, Jared, Boodle, Well]

---

Superstore

[Amy, Well, Glenn, Dina, Okay, Yeah, Jonah, Cheyenne, Hey, Lets, Garrett, Mateo, Frank, Go, Look, Yo, Oh]

---

Veep

[GARY, SELINA, Hallowes, Maam, Great, Kagans, Meyer, Codeine, Madam, ANNA, Dan, Veep, Frank, Carol, Reeves, Mike, Jonah, Tom, retard]

---

Medium revised version 2 list. Using .ents for inner loop instead of for tok in sent: (still no Jonah ☹☹):

30 rock

[LISA, Kenneth, FRANK, Josh, PETE, Gary, JACK, SARIE, Jenna, LAWRENCE, Donaghy]

---

B99

[CHARLES, TERRY, Holt, Gina, Rosa, Rihanna, Peralta, Jake, Boyle, Kane, Santiago, Morgenthau, Detective, Diaz]

---

Its always sunny

[DENNIS, CHARLIE, Mac, Dee, TERRELL, Yo, KAREN, Janelle]

---

Mindy project

[Mindy, Castellano, Danny, Tom, Gwen, Dennis, Carl, Springsteen]

---

Parks and rec

[LESLIE, Ron, Tom, Barry, Mayor, Parks, Ann, Aubrey, Mark]

---

The office

[Michael, Jim, Oscar, Scott, Dwight, Ryan, Roy]

---

Scrubs

[JAMES, ELLIOT, JEFFERY, Bambi, Carla, Dorian, Reid, Kelso, Pratt, Burski, Turk]

---

Silicon valley

[Steve, Thomas, Gregory, MONICA, ROLAND, LANGDON, Gavin, Jared, Belson, Peter, Boodle]

---

Superstore

[Amy, Glenn, Dina, Garrett, Frank, Mateo]

---

Veep

[GARY, SELINA, Maam, Hallowes, Dan, Veep, Frank, Reeves, Mike, Anna, Tom, Meyer]

---

Medium version 3, still no Jonah (not even in Veep this time):

30 rock

['LISA', 'Kenneth', 'FRANK', 'Josh', 'PETE', 'Gary', 'JACK', 'SARIE', 'Jenna', 'LAWRENCE', 'Donaghy']

---

B99

['CHARLES', 'TERRY', 'Holt', 'Gina', 'Rosa', 'Rihanna', 'Peralta', 'Jake', 'Boyle', 'Kane', 'Santiago', 'Morgenthau', 'Detective', 'Diaz']

---

Its always sunny

['DENNIS', 'CHARLIE', 'Mac', 'Dee', 'TERRELL', 'Yo', 'KAREN', 'Janelle']

---

Mindy project

['Mindy', 'Castellano', 'Danny', 'Tom', 'Gwen', 'Dennis', 'Carl', 'Springsteen']

---

Parks and rec

['LESLIE', 'Ron', 'Tom', 'Barry', 'Mayor', 'Parks', 'Ann', 'Aubrey', 'Mark']

---

The office

['Michael', 'Jim', 'Oscar', 'Scott', 'Dwight', 'Ryan', 'Roy']

---

Scrubs

['JAMES', 'ELLIOT', 'JEFFERY', 'Bambi', 'Carla', 'Dorian', 'Reid', 'Kelso', 'Pratt', 'Burski', 'Turk']

---

Silicon valley

['Steve', 'Thomas', 'Gregory', 'MONICA', 'ROLAND', 'LANGDON', 'Gavin', 'Jared', 'Belson', 'Peter', 'Boodle']

---

Superstore

['Amy', 'Glenn', 'Dina', 'Garrett', 'Frank', 'Mateo']

---

Veep

['GARY', 'SELINA', 'Maam', 'Hallowes', 'Dan', 'Veep', 'Frank', 'Reeves', 'Mike', 'Anna', 'Tom', 'Meyer']

1. Now focus on dialogue extraction

Final list after stop word removal (we have Jonah for SS!! But lost Mark from Parks and Rec ☹ )

30 rock

['LISA', 'KENNETH', 'PETE', 'FRANK', 'JOSH', 'Gary', 'JACK', 'SARIE', 'Jenna', 'Donaghy', 'Lawrence']

---

B99

['JAKE', 'Morgenthau', 'CHARLES', 'TERRY', 'Holt', 'Gina', 'Rihanna', 'Peralta', 'Rosa', 'Kane', 'Clay', 'Santiago', 'Boyle', 'Diaz', 'Detective']

---

Its always sunny

['DENNIS', 'CHARLIE', 'Mac', 'TERRELL', 'Dee', 'Yo', 'KAREN', 'Janelle']

---

Mindy project

['Mindy', 'Castellano', 'Danny', 'Tom', 'Gwen', 'Dennis', 'Carl', 'Springsteen']

---

Parks and rec

['LESLIE', 'Ron', 'Tom', 'Barry', 'ANN', 'ANDY', 'Mayor', 'Parks', 'Aubrey', 'Pawnee']

---

The office

['Michael', 'Jim', 'Pam', 'Oscar', 'Scott', 'Dwight', 'Stanley', 'Roy', 'Ryan']

---

Scrubs

['JAMES', 'ELLIOT', 'JEFFERY', 'Bambi', 'Carla', 'Dorian', 'Reid', 'Kelso', 'Pratt', 'Burski', 'ROBERTS']

---

Silicon valley

['Thomas', 'Steve', 'Boodle', 'Gregory', 'MONICA', 'ROLAND', 'LANGDON', 'JARED', 'Gavin', 'Belson']

---

Superstore

['Amy', 'Glenn', 'Dina', 'Mateo', 'Garrett', 'Frank', 'Jonah']

---

Veep

['GARY', 'SELINA', 'Hallowes', 'Maam', 'Veep', 'DORSEY', 'DAN', 'Frank', 'Carol', 'Reeves', 'MIKE', 'Jonah', 'Anna', 'Tom', 'Meyer']

---

Above approach after removing stop words takes about 53 seconds.

Moving fwd with extraction rn.

1. Maybe later can do something with other identity types (e.g., ORG). Or can compare entity-wise counts/graphs across scripts. Some stories/jokes may be extremely situational (situational comedy) and some may be based on pop culture/broad frame of reference (punchline-based humor)

Generally character followed by number means non essential character. E.g. Mom 1, Customer 2.

**DOCUMENT SIMILARITY W. SPACY**

**Currently removed all chars except space and colon.**

**Male and female character dialogue distribution??**

NER results without retaining colons:

30 rock

['LISA', 'KENNETH', 'PETE', 'FRANK', 'JOSH', 'Gary', 'JACK', 'SARIE', 'Jenna', 'Donaghy', 'Lawrence']

---

B99

['JAKE', 'Morgenthau', 'CHARLES', 'TERRY', 'Holt', 'Gina', 'Rihanna', 'Peralta', 'Rosa', 'Kane', 'Clay', 'Santiago', 'Boyle', 'Diaz', 'Detective']

---

Its always sunny

['DENNIS', 'CHARLIE', 'Mac', 'TERRELL', 'Dee', 'Yo', 'KAREN', 'Janelle']

---

Mindy project

['Mindy', 'Castellano', 'Danny', 'Tom', 'Gwen', 'Dennis', 'Carl', 'Springsteen']

---

Parks and rec

['LESLIE', 'Ron', 'Tom', 'Barry', 'ANN', 'ANDY', 'Mayor', 'Parks', 'Aubrey', 'Pawnee']

---

The office

['Michael', 'Jim', 'Pam', 'Oscar', 'Scott', 'Dwight', 'Stanley', 'Roy', 'Ryan']

---

Scrubs

['JAMESIm', 'JAMES', 'ELLIOT', 'Turk', 'JEFFERY', 'Bambi', 'Carla', 'Dorian', 'Reid', 'Kelso', 'Pratt', 'Burski', 'ROBERTS']

---

Silicon valley

['Thomas', 'Steve', 'Boodle', 'Gregory', 'MONICA', 'ROLAND', 'LANGDON', 'JARED', 'Gavin', 'Belson']

---

Superstore

['Amy', 'Glenn', 'Dina', 'Mateo', 'Garrett', 'Frank', 'Jonah']

---

Veep

['GARY', 'SELINA', 'Hallowes', 'Maam', 'Veep', 'DORSEY', 'DAN', 'Frank', 'Carol', 'Reeves', 'MIKE', 'Jonah', 'Anna', 'Tom', 'Meyer']

**#print(match\_str)**

**#span1=doc.char\_span(start,match\_str.index(' '))**

**#span1=match\_str[0:match\_str.index(" ")]**

**#print(start,start1)**

**#span2=doc.char\_span(match\_str.index(' ')+1,end)**

**#span2=match\_str[match\_str.index(' ')+1:len(match\_str)]**

**#end2=len(match\_str)-match\_str.index(' ')**

**#end2=start1+len(span1)-match\_str.index(' ')**

**#end1=(end-start1)**

**#print(end1,(end+1))**

**#print(span2)**

**#print(my\_tok)**

[(106, 108, 'Michael Scott'), (112, 114, 'Regional Manager'), (115, 117, 'Dunder Mifflin'), (117, 119, 'Paper Products'), (260, 262, 'Pam Beesly'), (431, 433, 'Spencer Gifts'), (602, 604, 'Jan Levinson'), (613, 615, 'Hillary Rodham'), (636, 639, 'Because Im'), (1081, 1083, 'Todd Packer'), (1386, 1388, 'Ryan Howard'), (1406, 1408, 'Moe Howard'), (1409, 1411, 'Three Stooges'), (1770, 1772, 'Steve Austin'), (1899, 1901, 'Bob Hope'), (1902, 1904, 'Abraham Lincoln'), (1968, 1971, 'Im Assistant'), (2106, 2108, 'And Jan'), (2511, 2513, 'Dwight Schrute'), (2549, 2551, 'Assistant Regional'), (2698, 2700, 'Its OK'), (2716, 2718, 'Sheriffs Deputy'), (3370, 3372, 'The Jamie'), (3372, 3374, 'Kennedy Experiment'), (3548, 3550, 'Postit Notes')]

Michael Scott PERSON

Regional Manager PERSON

Dunder Mifflin PERSON

Paper Products PERSON

Pam Beesly PERSON

Spencer Gifts PERSON

Jan Levinson PERSON

Hillary Rodham PERSON

Because Im PERSON

Todd Packer PERSON

Ryan Howard PERSON

Moe Howard PERSON

Three Stooges PERSON

Steve Austin PERSON

Bob Hope PERSON

Abraham Lincoln PERSON

Im Assistant PERSON

And Jan PERSON

Dwight Schrute PERSON

Assistant Regional PERSON

Its OK PERSON

Sheriffs Deputy PERSON

The Jamie PERSON

Kennedy Experiment PERSON

Postit Notes PERSON