Sitcom topic analysis :

Friends and *Brooklyn Nine-Nine*

LDA-H503

Introduction to Methods in Digital Humanities

Pia Julia Malinen

University of Helsinki

Introduction

For this project, I was interested in finding out what topics are dealt with in popular American television situational comedies to see how popular media may, to an extent, reflect what issues we (referring to us as a Western audience) value and find prominent in current discussions. As this research could indeed be nearly unlimited, I decided to limit the scope to two television shows: *Friends* and *Brooklyn Nine-Nine*. The reason for this is that I am familiar with both of them (I recognize this may cause a certain degree of bias which I have attempted to eliminate as much as possible), they are both extremely popular, and they are set 10 years apart which allows us to better observe any topical differences. My humanities research questions are as follows:

1. What are the most frequent topics in both situational comedies? How have they changed over the years and why?
2. What do these topics and changes tell us about current topics within popular media? How is this connected to (American) society in the 1990s-2010s on a larger scale?

This project is structured in such a way that I first present what kind of data I used, then I go through each step of the pipeline (1-5) by explaining what I did, presenting my results, and discussing the issues and limitations I faced. In my final step I will discuss the results from a humanities perspective, after which I will say what I would have done differently.

Data

The data I used began from the episode transcripts. As the original transcripts are not available for public use, I have used fan-generated transcripts that I found online. To verify their validity, I cross-referenced them across multiple fan-generated transcript websites and rewatched some episodes to ensure that they were accurate. The transcripts I used for *Friends* were from https://fangj.github.io/friends/ [1]. The transcripts I used for *Brooklyn Nine-Nine*

[1] Please note that for the Season 1 Episode 1 transcript, there were some sentences in blue which are supposedly previously unseen parts of the episode. These sentences did not appear in the versions of the episode that I watched, but I am uncertain whether or not the online versions of the episode have been edited and the one with the blue text is the original one. Therefore, I kept the blue parts in for the pilot.

were from https://www.springfieldspringfield.co.uk/episode_scripts.php?tv-show=brooklyn-nine . It should be noted that for these transcripts, there were no character names used (unlike in the *Friends* transcripts), but I chose this website regardless due to the fact that I would not be counting in the names as topics later on regardless. I considered using opensubtitles.org for downloading the subtitles for the episodes, but based on online reviews this website has some adware and malware issues, so I therefore chose not to risk it. The sources I ended up choosing were the best possible ones as for many other websites (such as https://transcripts.foreverdreaming.org/viewforum.php?f=429), the text would mold itself into one large chunk, meaning each word, sentence, and line would have had to be manually separated.

It should be noted that the data pools are not equal. *Friends* has 10 seasons to draw from whereas *Brooklyn Nine-Nine* has 6 (though there will be more in the upcoming years). Thus the results will look slightly different, as *Friends* has a wider range of possibilities to draw from.

Data processing

The process for tackling the data went as follows:

1. **Word**

I copy & pasted the first episode of every season from *Friends* (10 episodes altogether) and *Brooklyn Nine-Nine* (6 episodes altogether) into their own Word files and saved them as .docx files. The reason I used the first episode of each season as opposed to random selection was due to the fact that the opening episode of television shows is often used to foster excitement amongst the audience and give a glimpse of what is to come within the upcoming season. Granted, this is not always the case, but for the sake of not overcomplicating the data, I found it to be most useful to approach it like this.

2. **Annif**

I then copy & pasted each episode individually into the Annif web user interface. Please note that I used the web interface as opposed to the downloadable tool which requires more time and programming knowledge due to the facts that I: a) do not feel confident or experienced enough with programming to adequately be able to use the more advanced tool, and b) only wanted to get a first look and overview of topics and subject and thus did not want to spend the majority of the project only attempting to do this step.

Annif limitations:

I am aware that this simplified output is limited as it is not designed for television show transcripts (and therefore will not show the most applicable results to these but will instead show the results as if it were reading it as a newspaper or book) and it does not show the percentages of the results displayed (only provides a visual implication in the form of horizontal bar). This means that you are required to be critical towards the output and make some generalizations and conclusions yourself while trying to avoid bias as much as possible.

Annif process:

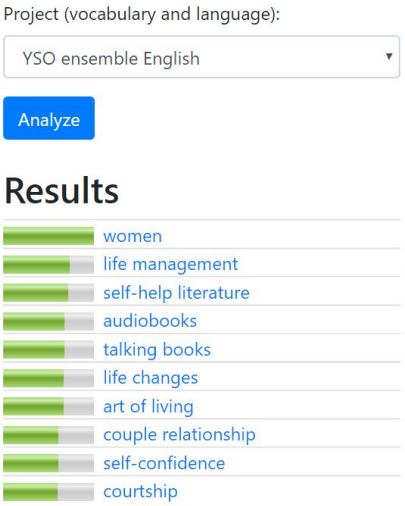
I copy & pasted my transcript into the web interface tool and began to observe what the results yielded based on the selected project output from the list of options (ie TF-IDF, Maui, fastText, YSO). I opted for the YSO English (yleinen suomalainen ontologia, eng 'general Finnish ontology') output, as this yielded the most clear and applicable categorizations and entities for the type of text I was using. According to their website,[2] YSO's primary focus is to categorize chiefly general concepts into one or more thematic groups. However, it should be noted that this categorization is biased due to the fact that it has been formed from Finnish cultural districts, and therefore the distinctions and outputs will be presented from this lens. I considered using the TF-IDF (term frequencies and inverse document frequencies) output to eliminate this bias as much as possible. With this output, the algorithm seeks to find and prioritize the words that are most crucial to a corpus, creating subject matters from these hierarchized words. While many of the subject matters were primarily similar to those that

[2] https://finto.fi/yso/fi/

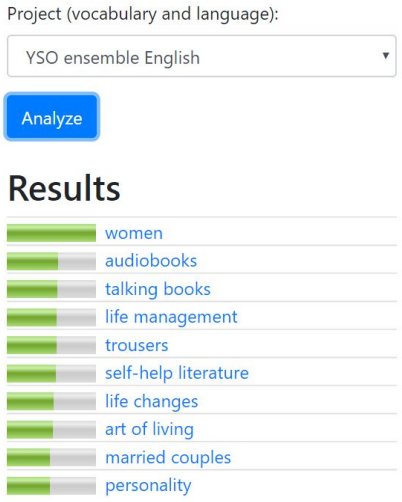
the YSO output generated, I found that with TF-IDF English there appeared to be more irrelevant topics as the algorithm appeared to be processing the data more as a more traditional, book-like text (with the most appearing subjects being ‘talking books’ and ‘audiobooks’ due to the structure of the dialogue within the text). Therefore, with TF-IDF, the central problem was the interface’s issue with not recognizing the text as a transcript and seeking subjects *within* it, but instead seeing the text’s structure itself *as* the primary subject. Due to this I stuck with YSO which was not perfect either but gave me more to work from.

Annif YSO results: *Friends*

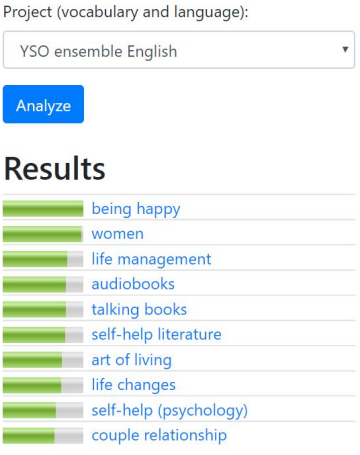
Season 1 episode 1:



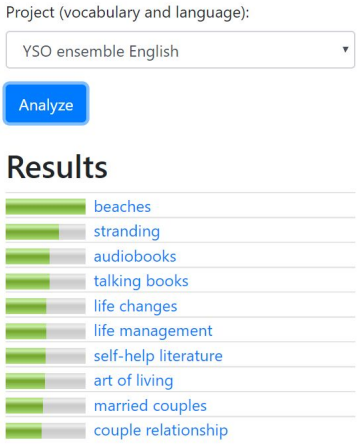
Season 2 episode 1:



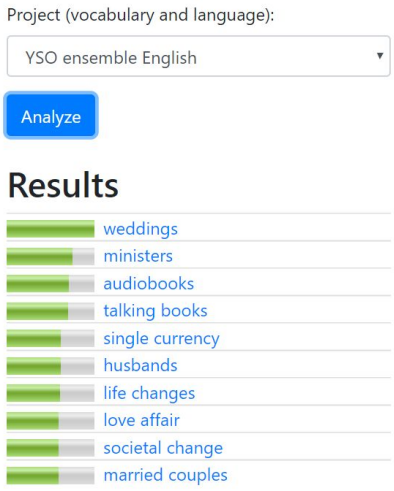
Season 3 episode 1:



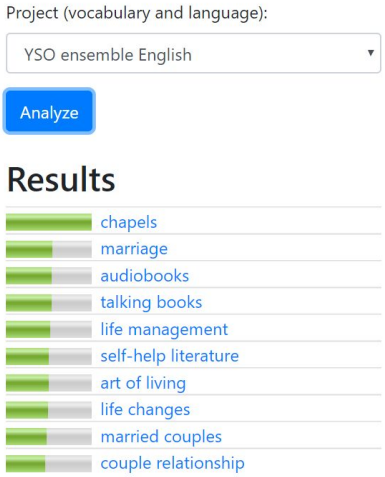
Season 4 episode 1:



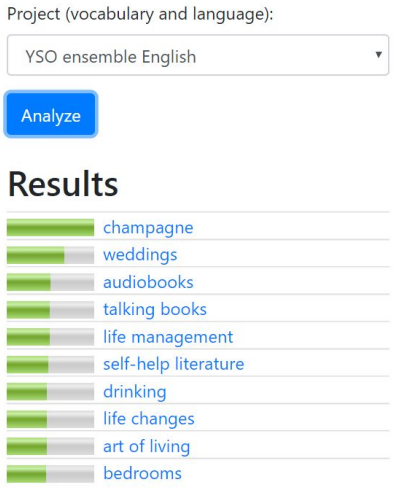
Season 5 episode 1:



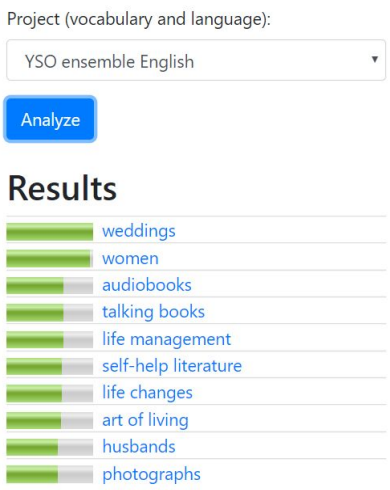
Season 6 episode 1:



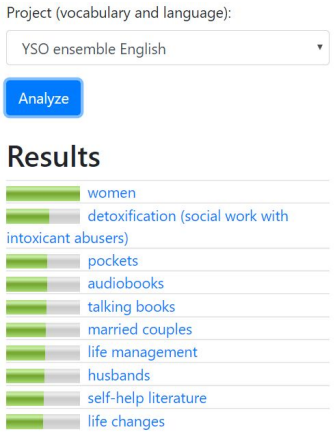
Season 7 episode 1:



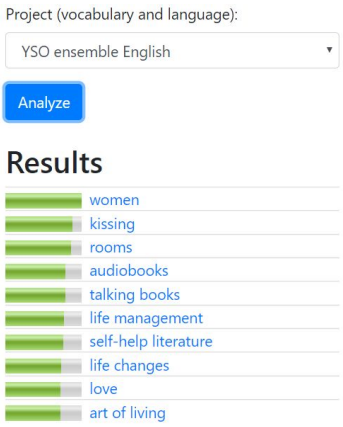
Season 8 episode 1:



Season 9 episode 1:



Season 10 episode 1:



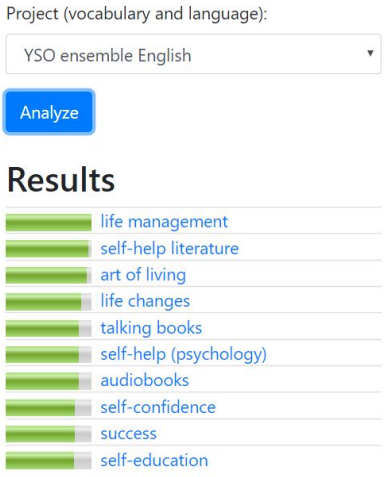
Based on Annif's results, the primary words that at first glance appeared to show up in most *Friends* seasons as topic categories were:

1. Women
2. Life management
3. Art of living
4. Couple
5. Life changes
6. Love

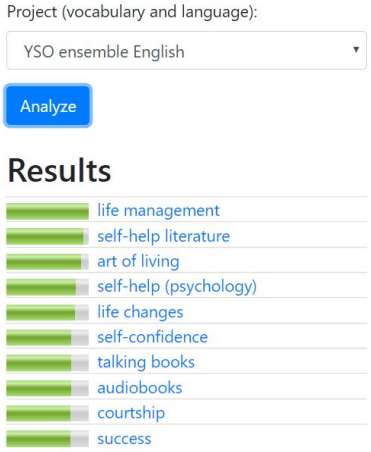
If wanting to eliminate more bias, however, you could choose not to observe what words appear to show up at first glance, as this may potentially skew the way in which you process your textual data later on.

Annif YSO results : *Brooklyn Nine-Nine*

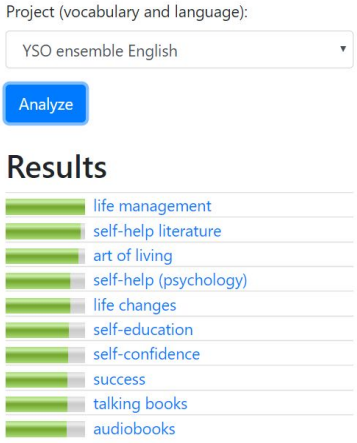
Season 1 episode 1:



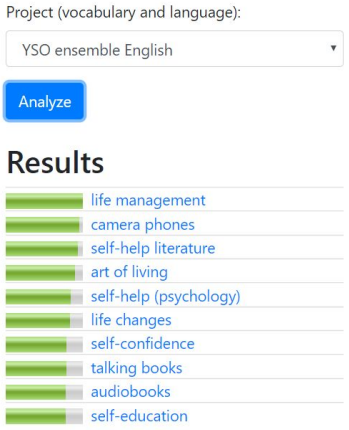
Season 2 episode 1:



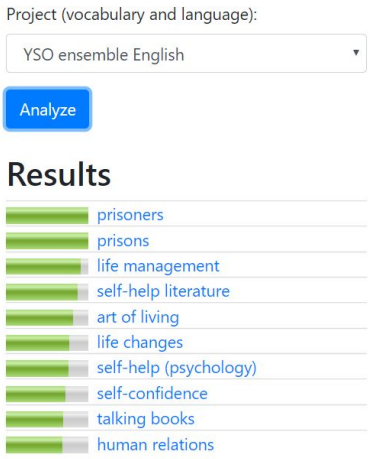
Season 3 episode 1:



Season 4 episode 1:



Season 5 episode 1:



Season 6 episode 1:

Based on Annif's results, the primary words that at first glance appeared to show up in most *Brooklyn Nine-Nine* seasons as categories are:

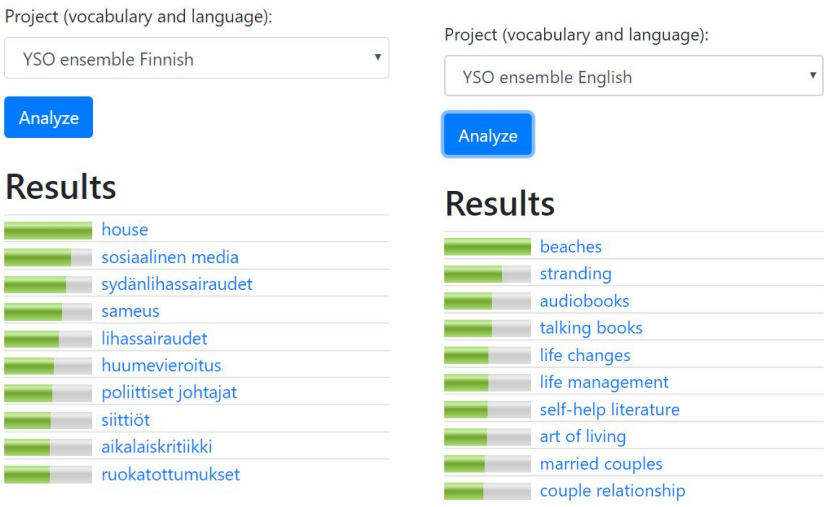
1. Life management
2. Life changes
3. Success
4. Art of living
5. Self-confidence

Please note that 'audiobooks,' 'talking books', and 'self-help literature' have been omitted from both frequent topic lists due to their lack of relevance and possible issue with the processing interface. Annif was originally built for subject indexing topics from metadata that consists of more formal texts such as scanned books and newspapers.[3] Therefore, it is crucial to note that this is not the most ideal interface for transcripts as the algorithm will likely suggest more irrelevant topics, which is why it is important to be critical towards the results.

To verify that this omission was justified, I rewatched some of the episodes to ensure that the primary topic was not relating to the omitted words. It should also be noted that when running the text through Annif and opting for Finnish YSO output, the suggested topics

[3] Suominen, Osma. "Annif: DIY Automated Subject Indexing Using Multiple Algorithms." LIBER Quarterly, July 29, 2019. <https://www.liberquarterly.eu/article/10.18352/lq.10285/.>

varied greatly. Here is a comparison between the Finnish YSO and English YSO output for *Friends* Season 4 episode 1:



Due to this vast difference of topicality, it is imperative to be critical of the output. However, a large reason that the Finnish YSO output may vary so greatly could be due to the fact that the algorithm is attempting to find subjects in Finnish from an English text which thus leads to some odd results. Therefore, if wanting to get more accurate results in Finnish, one would have to use transcripts that have been translated into Finnish as the input if wishing to compare and contrast between the two languages. The issue with this is not only the fact that you are adding an additional (potentially useless) layer to the process, but you are also limiting reproducibility by assuming those reproducing the project speak Finnish as well.

3. **Tables**

I then created tables for a visual indication of topics and to have a side-by-side comparison of the occurring topics, omitting neutral or useless words and phrases as mentioned before. The topics are displayed in descending order (the most prominent topics at the top) in a similar way as in the Annif output's coloured in bars. I color-coded each panel based on subject similarity which I will then later use to create my primary subject categories. For now I separated similarities based on the following colour logic[4] :

[4] I recognize that this logic is biased as it stems from me and therefore cannot be wholly subjective. To avoid this personal (human) bias, programming language would be required.

Pink = gendered elements

Blue = relationship elements

Green = introspective and life skills elements

Yellow = possible links to other categories but primarily unclassifiable

Table 1: topic frequency in *Friends*

1994 (S1E1)	1995 (S2E1)	1996 (S3E1)	1997 (S4E1)	1998 (S5E1)	1999 (S6E1)	2000 (S7E1)	2001 (S8E1)	2002 (S9E1)	2003 (S10E1)
Women	Women	Being happy	Stranding	Weddings	Chapels	Weddings	Weddings	Women	Women
Life management	Life management	Women	Life changes	Ministers	Marriage	Life management	Women	Detoxification	Kissing
Life changes	Life changes	Life management	Life management	Single currency	Life management	Drinking	Life management	Married couples	Life management
Art of living	Art of living	Art of living	Art of living	Husbands	Art of living	Life changes	Life changes	Life management	Life changes
Couple relationship	Married couples	Life changes	Married couples	Life changes	Life changes	Art of living	Art of living	Husbands	Love
Self-confidence	Personality	Self-help	Couple relationship	Love affair	Married couples		Husbands	Life changes	Art of living
Courtship		Couple relationship		Societal change	Couple relationship				
				Married couples					

I cut out object terms such as ‘pockets’, ‘rooms’ and ‘trousers’ that I could not categorize into subjects due to the nature of their neutrality.

I then did the same with *Brooklyn Nine-Nine.* However, I added a new colour, as a new topic had seemed to appear.

Purple = career elements[5]

Pink = gendered elements

Blue = relationship elements

Green = introspective and life skills elements

Yellow = possible links to other categories but primarily unclassifiable

Table 2 : topic frequency in *Brooklyn Nine-Nine*

2013 (S1E1)	2014 (S2E1)	2015 (S3E1)	2016 (S4E1)	2017 (S5E1)	2018 (S6E1)
Life management	Life management	Life management	Life management	Prisoners	Life management
Art of living	Art of living	Art of living	Camera phones*	Life management	Art of living
Life changes	Life changes	Life changes	Art of living	Art of living	Life changes
Self-confidence	Courtship	Self-confidence	Life changes	Life changes	Self-confidence
Success	Success	Success	Self-confidence	Self-confidence	Self-help
Self-education			Self-education	Human relations	Self-education
					Success

* Although 'camera phones' is an object category, I chose to keep it in due to the fact that it as an object can be linked to a certain moment in time which provides us with another level of decade-specific distinction and comparison to *Friends.*

Based on the two topic tables I created, I created 4 main topic categories:

1. Love & relationships
2. Career
3. Introspection (referring to the observation and inspection of the self)
4. Gender

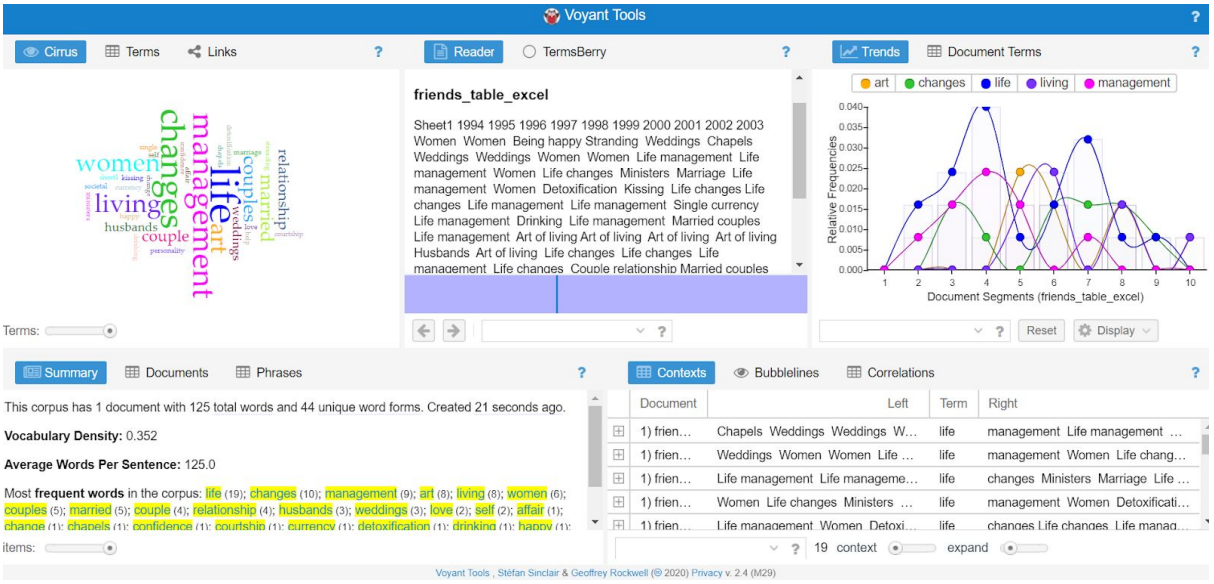
[5] It should be noted, however, that this is very subjective. The primary word underneath this topic is 'success', but as Annif does not clarify what type of success this is, it is necessary to create these conclusions yourself. I chose to draw these conclusions as I am familiar with the series and am aware of the emphasis on career success, but this is important to note for someone reproducing the project that in order to make these judgements through this methodology it is crucial to be very familiar with the show, which further fosters bias.

I recognize that drawing my own conclusions from the presented data creates bias. Python's NLP or LDA programme would have been more ideal to do this for me more categorically (and in a less biased way), but due to lack of skill I chose to do this manually, which will inevitably have its own issues.

4. **Voyant**

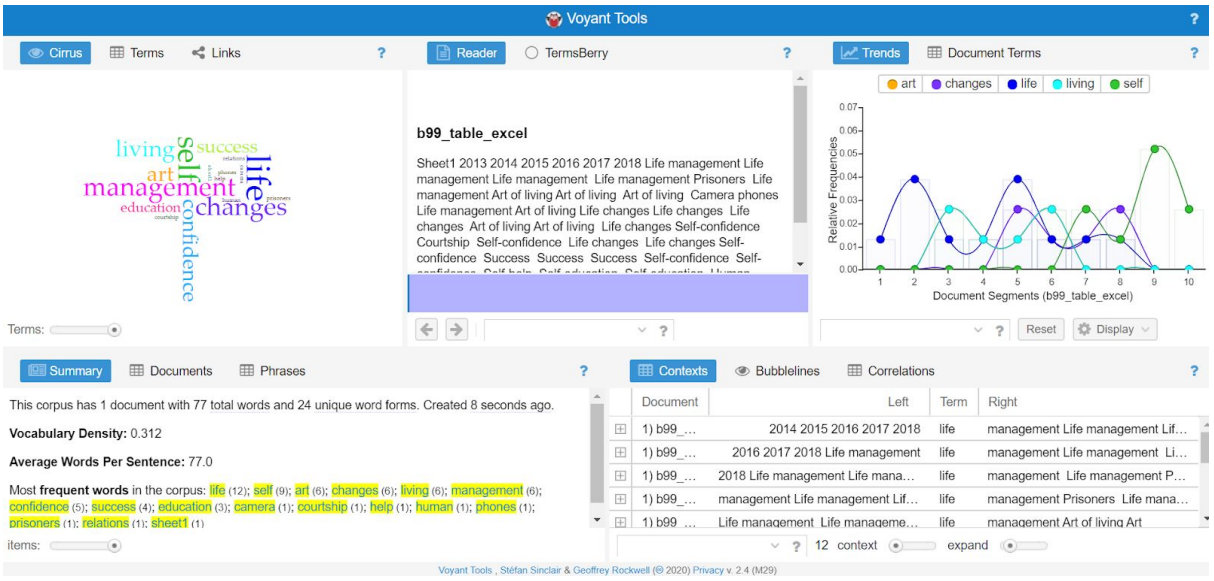
I then wanted to get a visual overview of the tables. I did not use this step before (ie copy & pasting the raw transcripts as they were) due to the fact that this would have had too many variables and yielded very messy results. Therefore, before getting to this step, it was crucial to clean up the text into more palatable terms for the tool. I copy & pasted the tables[6] into Excel and saved them as .xlsx files, which I then uploaded into Voyant's web-based text reading and analysis. Note that Voyant cuts out words that occur multiple times (such as 'the') to avoid categorizing them with more generalized topics.

Voyant results : *Friends* topic table visualized



[6] I attempted to export the transcript linguistic data into Excel via processing it as a txt file then converting into csv, but this did not really work. I found that copy & pasting was sufficient, so I went with that.

Voyant results : *Brooklyn Nine-Nine* topic table visualized



I wanted to display all the results that Voyant showed me, as using only a word cloud (top left) would have been fairly reductive. According to Voyant's analysis (and based on what we could draw from the tables as well), *Friends* deals with a wider range of topics (as predicted possibly due to its larger number of seasons). However, it should be noted that Voyant's data can be misleading, as for example the relative frequency chart (top right) for *Friends* only goes up to 0.04, whereas for *Brooklyn Nine-Nine* it goes up to 0.07. If comparing side-by-side, this is an example of a detail that may be easily missed and thus lead to false conclusions. Regardless, based on the visualized data above, a prominent theme in both sitcoms is 'life' and 'changes'; for *Friends* 'management' and 'women' also occupy a large role, whereas for *Brooklyn Nine-Nine* 'confidence' and 'self' appear more frequently.

5. **Math + RAW**

2014
Life management
Art of living
Life changes
Courtship
Success

In order to further visualize the data on RAW, I needed to convert the information into strings of data on Excel. I did this by simply counting the amount of times each topic category and year occurred in the tables above. For example, for *Brooklyn Nine-Nine*, we can observe 5 primary topic categories for the year 2014 (S2E1) in the table: life management, art of living, life changes, courtship, and success. Therefore we must divide 100% into the amount of topic categories to get how many percent each is worth. The calculation for this is:

$$100 \div \textit{number of categories in each panel} \times \textit{panel amount of particular category}$$

In this example, each individual panel for 2014 is worth 20% (100/5). As we can see there are 3 green panels, so we must multiply 20x3 = 60%. The other two categories are worth 20%. Therefore, for this example, the 'introspection' category is worth 60%, 'love and relationships' is 20%, and 'career' is 20%. I then did these calculations for each year and category and applied my results into Excel cells (one for each individual show and one for the results combined). I had to change the percentages to numbers (ie 0,37) and the years into a different format (ie 2014-01) so that they could be processed in RAW (the requirements for the types of data RAW will accept can be found on their website.[7]) I rounded the percentages to the nearest 0.01 percent for simplification.

Excel screenshots : *Friends* (left) and *Brooklyn Nine-Nine* (right)

topics split into years and percentages

I53

	A	B	C
1	Category	Year	Percentages
2	Love & rel	1994-01	0,29
3	Career	1994-01	0
4	Introspect	1994-01	0,57
5	Gender	1994-01	0,14
6	Other	1994-01	0
7	Love & rel	1995-01	0,17
8	Career	1995-01	0
9	Introspect	1995-01	0,67
10	Gender	1995-01	0,17
11	Other	1995-01	0
12	Love & rel	1996-01	0,14
13	Career	1996-01	0
14	Introspect	1996-01	0,71
15	Gender	1996-01	0,14
16	Other	1996-01	0
17	Love & rel	1997-01	0,33
18	Career	1997-01	0
19	Introspect	1997-01	0,5
20	Gender	1997-01	0
21	Other	1997-01	0,17

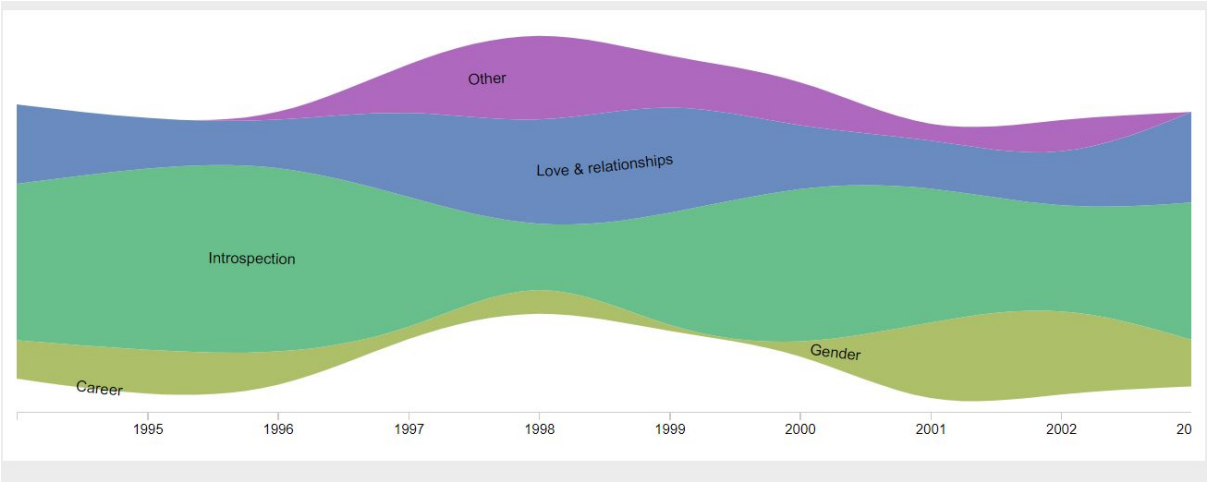
L14

	A	B	C
1	Category	Year	Percentages
2	Love & rel	2013-01	0,20
3	Career	2013-01	0,20
4	Introspect	2013-01	0,60
5	Gender	2013-01	0,00
6	Other	2013-01	0,00
7	Love & rel	2014-01	0,20
8	Career	2014-01	0,20
9	Introspect	2014-01	0,60
10	Gender	2014-01	0,00
11	Other	2014-01	0,00
12	Love & rel	2015-01	0,00
13	Career	2015-01	0,20
14	Introspect	2015-01	0,80
15	Gender	2015-01	0,00
16	Other	2015-01	0,00
17	Love & rel	2016-01	0,00
18	Career	2016-01	0,00
19	Introspect	2016-01	0,83
20	Gender	2016-01	0,00
21	Other	2016-01	0,17

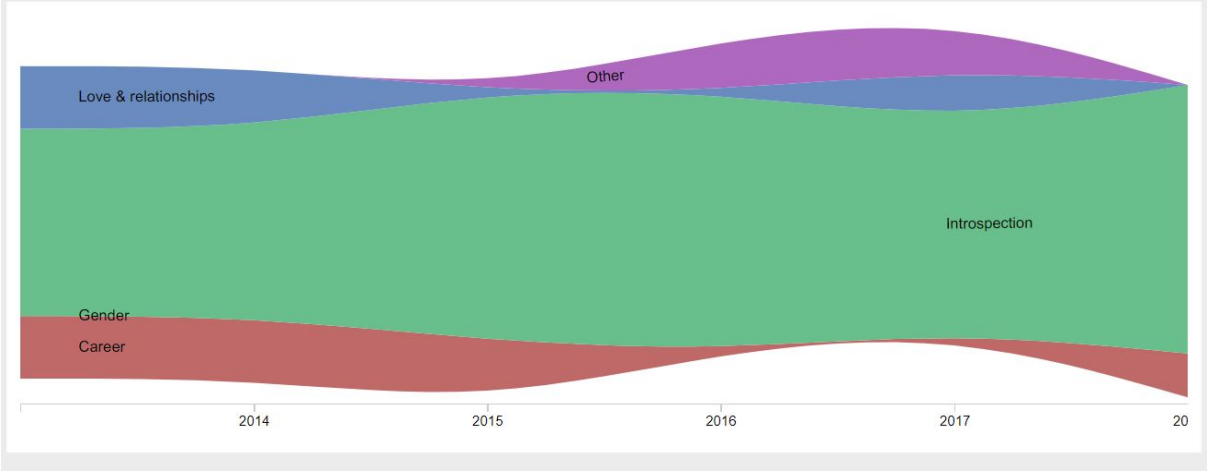
I then saved the data as a .xlsx file and uploaded it to RAW. My dimensions were 'Category' (string), 'Year' (date), and 'Percentages' (number). I was then able to use these dimensions to visualize a streamgraph of the results.

[7] https://rawgraphs.io/about

Friends topic streamgraph

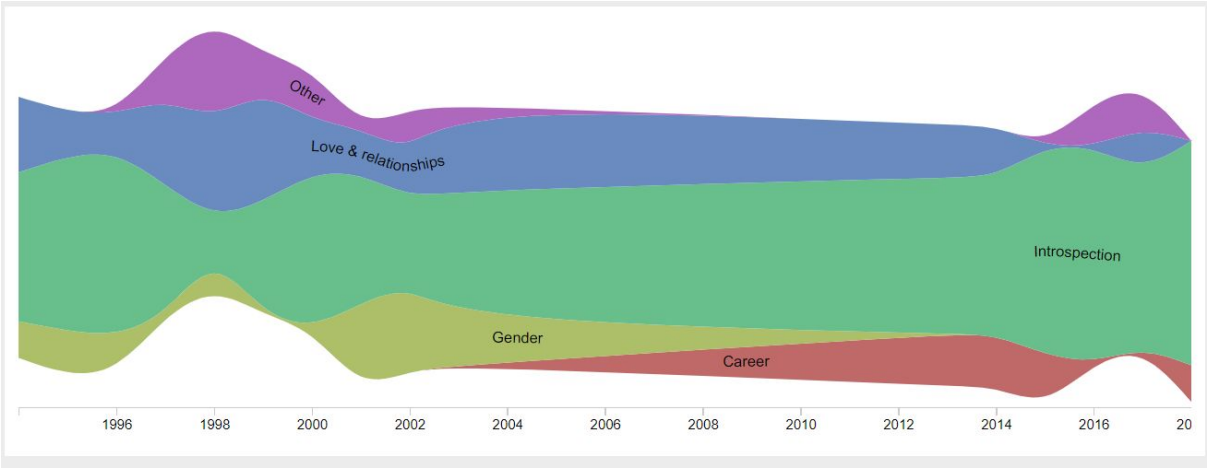


Brooklyn Nine-Nine topic streamgraph



The streamgraph is limited as it does not show the percentages, it includes the category 'Career' in *Friends* and 'Gender' in *Brooklyn Nine-Nine* although they did not occur in any of the data within these graphs, and it also cuts off the years 2003 and 2018. Nevertheless, I chose to go with a streamgraph due to the fact that it displays the prominence of the themes through a continuous amount of time.

Combining both into a single streamgraph



From the streamgraph above, we can observe which topics (out of the selected 5) have been the most prominent within these two sitcoms and when. It is important to note, however, that as *Friends* has more data, it occupies a larger chunk of the streamgraph (from a total of 16 seasons *Friends* is 10, meaning its topics occupy 63% of the graph). With this in mind, we can see that 'introspection' appears to have always been the most prominent theme, and as the themes of 'gender' and 'love and relationships' became less prominent, the theme of career became more prominent.

6. **Discussion of results**

So what exactly does this (limited) analysis show? Based on what these tools have visualized for us from these two shows, one could say that introspection has always been an important topic in sitcoms. It appears to be the backbone of the shows topics upon which the other subjects are formed. Considering that situational comedies are often based on character tropes and differences in their personalities, this comes as no surprise. However, what is interesting to consider is the increased emphasis on career alongside the decreased emphasis of love and relationships.

Something to consider when observing these results is the situation of America (and the world) during these eras. One could argue that the 1990s and early 2000s were a more peaceful time in the United States (at least until 2001) in comparison to the 2010s, which is

when I would argue that there is more visible political unrest, anxiety (about the planet and individual mental health for example), and loneliness with the rise of social media and digital technology. *Friends* was also filmed for a live studio audience, which is rarely done in American television anymore (except for talk shows). Because of this, the comedic rhythm had to be constructed in such a way that it would get an audibly live reaction, meaning there may have been less time for building up characters and jokes and instead the focus was on fast punchlines. With the rise of streaming services such as Netflix, audiences have the possibility to watch however many episodes they want in a row whenever they wish (which could previously only be done if you owned the DVD sets), and therefore writers have a better opportunity to build up characters, jokes, and situations. Hence why the introspective element appears larger.

The themes of the shows themselves are also indicative of their most common topics. *Friends* is a more feel-good, optimistic and romanticized presentation of life for young adults in New York in the 1990s, whereas *Brooklyn Nine-Nine* is a goofy, caricatured presentation of how un-serious a police precinct in New York can be. Therefore we could say that the topics themselves are as important as the way in which they are presented, which shows us that there has been a shift in tonality and the way in which we view sitcoms. It could be argued that between 1993-2003, sitcoms presented audiences with a larger range of topics and a faster pace of humor in order to provide them with as much feel-good subject material as they could fit within the 23-minute time frame. Since the 2010s, sitcoms have focused more on topics of the self and taken a slower pace so that the viewer is encouraged to watch more episodes in a row. A potential reason for this could be due to the shift of global and national discussions from more frivolous topics to ones that carry more weight as a reaction to the ever-changing world around us and what is being discussed on more platforms than ever before. With more outlets, news, and fear-mongering than ever, television sitcoms have the possibility and right to discuss surface-level topics, but most shows such as *Brooklyn Nine-Nine* choose not to do so. Show creators are aware of the fact that people's attention is a valuable commodity, and that cheap gags and laugh tracks no longer work for people the same way they used to. On a larger humanities scale, we could conclude that people have more interest in consuming more meaningful and socially conscious content due to the changing nature of the discussions occurring within the realm of popular culture. I am,

however, aware that on the flipside of this there is also an even greater plethora of "useless" content that people can consume (such as scripted reality television) to seek escapism as well, and that my conclusion is biased because of this and the fact that I was not observing or participating in humanities-related discussions in the 1990s.

What I would do differently in this project:

If I had the time to dedicate myself to improve my skills in programming, the best approach to this project would have been through NLP (natural language processing) programs, such as Python's NLTK (natural language toolkit) or something similar that yields a hypothesis and more focused results from more quantifiable data (for example csv). In order for the analysis to be more reliable, I would implement the following recommendations:

- choosing a larger pool of television shows
    → comparing and contrasting shows from different television networks, directors, writers, etc and taking all these factors into consideration
- choosing television shows that you are unfamiliar with (reducing bias)
- using more complex tools (such as Python mentioned above) to further reduce bias

Sources

Dalton, Mary M., and Laura R. Linder. *The Sitcom Reader: America Viewed and Skewed*. Albany, NY: State University of New York Press, 2005.

Garis, Mary Grace. "What Will Our Sitcoms Be Like In 2020?" Bustle, February 9, 2015. <https://www.bustle.com/articles/63052-evolution-of-the-television-sitcom-from-studying-1980-to-predicting-2020.>

Ihnat, Gwen. "How Friends Changed the Sitcom Landscape." TV Club, August 8, 2014. <https://tv.avclub.com/how-friends-changed-the-sitcom-landscape-1798271378.>

Mills, Brett. *The Sitcom*. Edinburgh: Edinburgh University Press, 2013.

Picone, Jack. "The Evolution Of The Sitcom: The Age of the Single Camera." Student Resources, April 26, 2018. <https://www.nyfa.edu/student-resources/evolution-sitcom-part-2/.>

Suominen, Osma. "Annif: DIY Automated Subject Indexing Using Multiple Algorithms." LIBER Quarterly, July 29, 2019. <https://www.liberquarterly.eu/article/10.18352/lq.10285/.>

Links:

Annif: http://annif.org/

Voyant: https://voyant-tools.org/

RAW: https://rawgraphs.io/