NLP Comparison Between Metamorphosis by Franz Kafka and Death of a Salesman by Arthur Miller

# Choosing the data:

These two stories were chosen based on the similarities in character development and story similarities. Metamorphosis is based on a salesman who physically transforms into a bug and is no longer useful to his family after years of providing them with a means of survival. Death of a Salesman explores the mental transformation of a salesman from a highly regarded bread-winner to a mentally unstable man who decides to take his life because his family no longer needs him. While these two stories have very different authors, I would like to explore the similarities and differences in their styles for telling stories with similar motifs.

## Forming the Corpora:

Metamorphosis was readily available from Project Gutenberg. However, Death of a Salesman was only available as a \*.pdf document that needed conversion to plain text. To convert Death of a Salesman to utf-8 format, I used the pdfminer function available from Python’s pypi library. This left me with two text files in similar format with which to work.

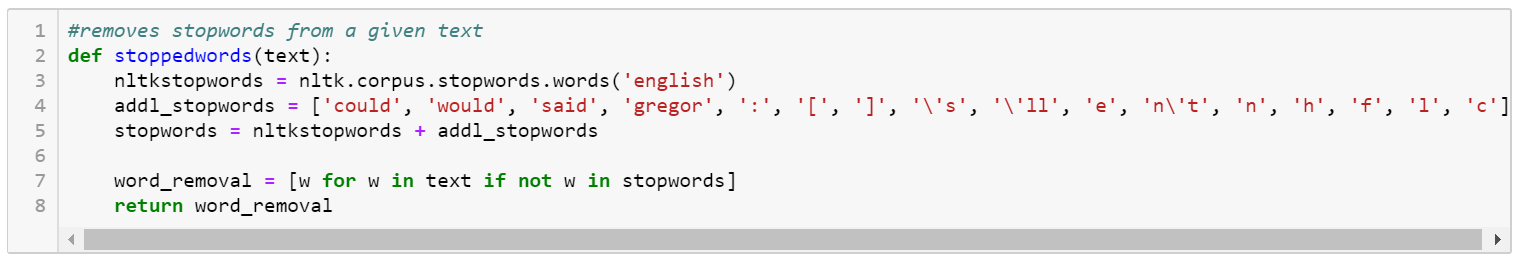
## Limitations to Text and Method:

Because Metamorphosis was originally written in German and translated into English, it may be difficult to directly attribute all language in this text to the author. However, after having read both texts, some similarities stand out that I would like to investigate further regarding word frequency, bigram similarity, and word usage. Also, this is a comparison of a play to a short story, so the formats are slightly different. This small difference means that Metamorphosis has a much denser word format as compared to Death of a Salesman, which has lots of text to describe sets and identify the speaker of each line in addition to the story’s text. Some editing in Python was done to minimize the effect of the metadata text found in Death of a Salesman.

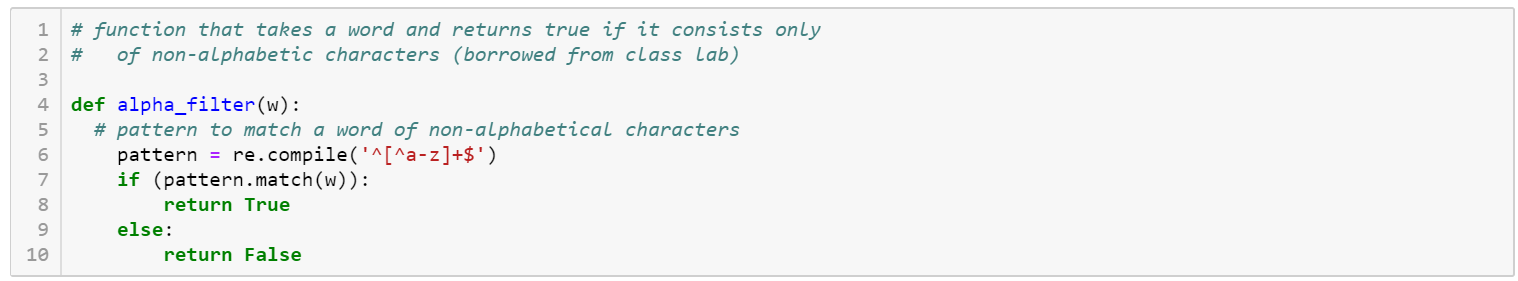
# Document Tokenization and Data Rules:

## Deciding which words to keep:

With regard to text lengths, Metamorphosis starts with 23970 words and Death of a Salesman starts with 41511. However, after accounting for punctuation and apostrophized letters, the two texts were minimized to 11017 and 17216 respectively. I arrived at these final numbers by including the nltk.corpus.stopwords function to remove common English words that carried little to no meaning for the context of either story. Along with these stop words, I added could, would, and said.



To remove punctuation, I incorporated a version of the regex pattern used in lab 2 called alpha\_filter. This took away lots of characters that were attached to metadata in the play’s text while also emphasizing the most important text in both works.



Additionally, the PorterStemmer function from nltk was used to help to gather words together that had similar meanings. Without lemmatization, many words appeared twice with multiple endings. I chose the Porter stemmer because overall it removed less of the words in most cases to where most words were still readable and recognizable from their stems.



## Tokenization

To tokenize, I followed the regex pattern that we used in lab3. This method was mostly used because Metamorphosis has lots of possessive nouns that point to relationships between the main character and others. This story also uses the word o’clock quite often. Time is a common motif in both stories, so I wanted to keep this word to make that comparison later on.

# Top 50 Tokens:

## Metamorphosis:

[['hi', 550], ['wa', 406], ['room', 110], ['even', 99], ["gregor'", 99], ['sister', 94], ['father', 93], ['door', 92], ['thi', 89], ['mother', 83], ['back', 80], ['one', 72], ['time', 66], ['onli', 61], ['look', 60], ['way', 60], ['open', 54], ['get', 52], ['go', 49], ['littl', 48], ['ani', 47], ['still', 45], ['first', 44], ['want', 43], ['like', 42], ['made', 40], ['see', 39], ['hand', 39], ['without', 39], ['head', 37], ['much', 37], ['chief', 37], ['make', 36], ['move', 36], ['turn', 35], ['come', 35], ['thing', 34], ['veri', 34], ['befor', 34], ['clerk', 34], ['samsa', 33], ['thought', 33], ['use', 33], ['seem', 31], ['well', 31], ['quit', 30], ['day', 30], ['soon', 30], ['let', 30], ['came', 30]]

## Death of a Salesman:

[['willi', 799], ['biff', 523], ['linda', 318], ['happi', 282], ['hi', 223], ['go', 208], ['wa', 156], ['get', 149], ['ben', 144], ['come', 138], ['know', 136], ['got', 129], ['charley', 124], ['see', 111], ['thi', 110], ['like', 106], ['well', 103], ['bernard', 103], ['take', 94], ['right', 92], ['look', 92], ['whi', 89], ['howard', 82], ['want', 81], ['oh', 76], ['talk', 76], ['him.', 74], ['back', 71], ['good', 71], ['man', 69], ['boy', 68], ['tell', 66], ['say', 66], ['make', 63], ['it.', 60], ['never', 60], ['think', 59], ['time', 59], ['ani', 57], ['put', 55], ['one', 54], ['thing', 53], ['woman', 52], ['stanley', 52], ['around', 49], ['let', 48], ['you.', 48], ['hous', 47], ['turn', 45], ['me.', 43]]

Most of the words in the word frequency lists look okay. Lemmatization was used to minimize repeated words with different suffixes and this fixed most duplicates. The only real problems caused by stemming words appears to be that some words are not recognizable. Also, the stemming of character names like Willy and Happy should have been omitted to retain the correct names for those people.

# Top 50 Bigrams:

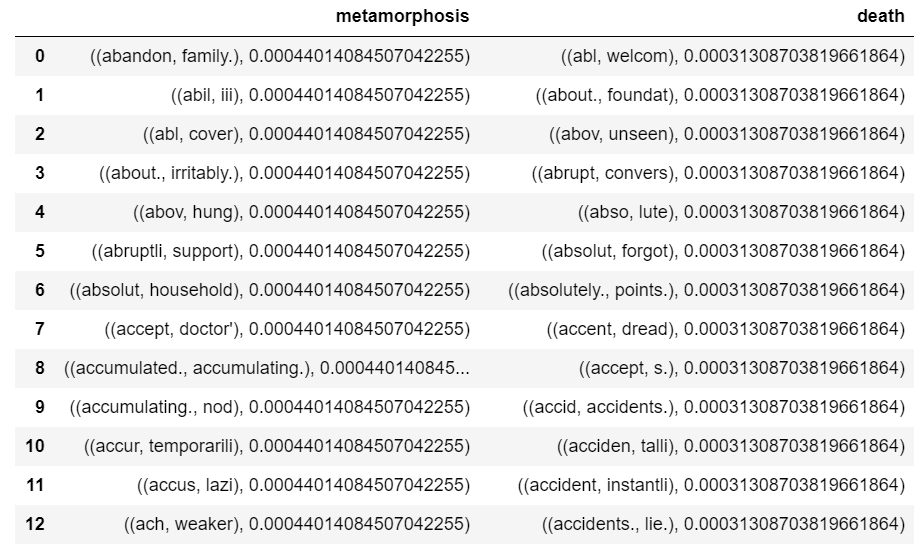
## Metamorphosis:

[(('abandon', 'family.'), 0.00044014084507042255), (('abil', 'iii'), 0.00044014084507042255), (('abl', 'cover'), 0.00044014084507042255), (('about.', 'irritably.'), 0.00044014084507042255), (('abov', 'hung'), 0.00044014084507042255), (('abruptli', 'support'), 0.00044014084507042255), (('absolut', 'household'), 0.00044014084507042255), (('accept', "doctor'"), 0.00044014084507042255), (('accumulated.', 'accumulating.'), 0.00044014084507042255), (('accumulating.', 'nod'), 0.00044014084507042255), (('accur', 'temporarili'), 0.00044014084507042255), (('accus', 'lazi'), 0.00044014084507042255), (('ach', 'weaker'), 0.00044014084507042255), (('achiev', 'distinguish'), 0.00044014084507042255), (('acquir', 'lock'), 0.00044014084507042255), (('across', 'balanc'), 0.00044014084507042255), (('act', 'violence.'), 0.00044014084507042255), (('action', 'action.'), 0.00044014084507042255), (('action.', 'ahead'), 0.00044014084507042255), (('actual', 'smell'), 0.00044014084507042255), (('ad', 'cope'), 0.00044014084507042255), (('add', 'selfish'), 0.00044014084507042255), (('address.', 'transport'), 0.00044014084507042255), (('adhes', 'tip'), 0.00044014084507042255), (('adjoin', 'room.'), 0.00044014084507042255), (('admit', 'fro'), 0.00044014084507042255), (("adult'", 'squar'), 0.00044014084507042255), (('advantage.', 'agre'), 0.00044014084507042255), (('advic', 'suffici'), 0.00044014084507042255), (('advis', 'variou'), 0.00044014084507042255), (('affair', 'apprentic'), 0.00044014084507042255), (('affect', 'return.'), 0.00044014084507042255), (('afraid', 'carri'), 0.00044014084507042255), (('afterward', 'fulli'), 0.00044014084507042255), (('afterwards.', 'lain'), 0.00044014084507042255), (('again.', 'gregor.'), 0.00044014084507042255), (('age', 'enthusiast'), 0.00044014084507042255), (('aggriev', 'implor'), 0.00044014084507042255), (('agit', 'bedroom'), 0.00044014084507042255), (('ago', 'gone'), 0.00044014084507042255), (('ago.', 'man'), 0.00044014084507042255), (('agre', 'spokesman'), 0.00044014084507042255), (('ah', 'angri'), 0.00044014084507042255), (('ahead', 'something.'), 0.00044014084507042255), (('ahead.', 'separ'), 0.00044014084507042255), (('aim', 'another.'), 0.00044014084507042255), (('air', 'occur'), 0.00044014084507042255), (('alarm', 'clock'), 0.00044014084507042255), (('alarm.', 'risked.'), 0.00044014084507042255), (('albeit', 'tremor'), 0.00044014084507042255)]

## Death of a Salesman:

[(('abl', 'welcom'), 0.00031308703819661864), (('about.', 'foundat'), 0.00031308703819661864), (('abov', 'unseen'), 0.00031308703819661864), (('abrupt', 'convers'), 0.00031308703819661864), (('abso', 'lute'), 0.00031308703819661864), (('absolut', 'forgot'), 0.00031308703819661864), (('absolutely.', 'points.'), 0.00031308703819661864), (('accent', 'dread'), 0.00031308703819661864), (('accept', 's.'), 0.00031308703819661864), (('accid', 'accidents.'), 0.00031308703819661864), (('acciden', 'talli'), 0.00031308703819661864), (('accident', 'instantli'), 0.00031308703819661864), (('accidents.', 'lie.'), 0.00031308703819661864), (('accommod', 'old'), 0.00031308703819661864), (('accompani', 'speech.'), 0.00031308703819661864), (('accomplish', 'something.'), 0.00031308703819661864), (('accomplishment.', 'reconstruct'), 0.00031308703819661864), (('account', 'inside.'), 0.00031308703819661864), (('accountant.', 'ington'), 0.00031308703819661864), (('accounts.', 'self.'), 0.00031308703819661864), (('accus', 'sult'), 0.00031308703819661864), (('ace', 'cards.'), 0.00031308703819661864), (('aces.', 'asham'), 0.00031308703819661864), (('across', 'huh'), 0.00031308703819661864), (('act', 'one'), 0.00031308703819661864), (('action', 'take'), 0.00031308703819661864), (('activ', 'count'), 0.00031308703819661864), (('actor', 'observ'), 0.00031308703819661864), (('actual', 'enough'), 0.00031308703819661864), (('ad', 'mire'), 0.00031308703819661864), (('add', 'consid'), 0.00031308703819661864), (('address', 'physic'), 0.00031308703819661864), (('address.', 'move.'), 0.00031308703819661864), (('addressing.', 'rag'), 0.00031308703819661864), (('admir', 'though'), 0.00031308703819661864), (('admiration.', 'shame'), 0.00031308703819661864), (('admit', 'per'), 0.00031308703819661864), (('adonises.', 'anc'), 0.00031308703819661864), (('advanc', 'premium.'), 0.00031308703819661864), (('advantag', 'maybe.'), 0.00031308703819661864), (('adventur', 'streak'), 0.00031308703819661864), (('advertis', 'gener'), 0.00031308703819661864), (('advic', 'advis'), 0.00031308703819661864), (('advice.', 'pennant'), 0.00031308703819661864), (('advis', 'teacher'), 0.00031308703819661864), (('af', 'ﬁrmativ'), 0.00031308703819661864), (('affec', 'tionately.'), 0.00031308703819661864), (('affect', 'known'), 0.00031308703819661864), (('afraid', 'storm'), 0.00031308703819661864), (('afraid.', 'implacably.'), 0.00031308703819661864)]

The bigrams list for each text has the same scores. I couldn’t figure out how to fix this to get better results. These results are also much more readable in a table view. Unfortunately, the full table wouldn’t fit here. Here are the top 12 scores in a table view:



Having the same score for each set of bigrams available makes it impossible to figure out which pairs are used most frequently. However, even with these two bigrams lists, we can see that there’s a lot of similarity in verbiage used by these two authors. Lots of negative sounding words like accuse, accident, afraid, and violence appear in both lists.

# Top 50 Bigrams by Mutual Information Score (PMI) [filter = 5]:

Metamorphosis:

[(('mrs.', 'samsa'), 9.50454816917804), (('chief', 'clerk'), 9.339488922907545), (('mr.', 'samsa'), 9.302914308008393), (('three', 'gentlemen'), 9.188879064762363), (('o', "'"), 8.96397978781534), (("'", 'll'), 8.963979787815338), (("'", 've'), 8.963979787815337), (("'", 're'), 8.83844890573148), (('some', 'kind'), 8.74158736647889), (('two', 'women'), 8.20909228565187), (('straight', 'away'), 8.156624865757735), (('little', 'legs'), 8.13390478925765), (('crawl', 'about'), 8.063515461366253), (("'", 'clock'), 7.963979787815337), (('each', 'other'), 7.946906274456399), (('my', 'parents'), 7.834696770870371), (('no', 'longer'), 7.711457700294506), (('living', 'room'), 7.291554445843843), (('other', 'side'), 7.227014193649131), (('must', 'have'), 7.096981919330334), (('at', 'least'), 7.010559328504375), (('even', 'though'), 6.794054786373026), (('we', "'"), 6.771334709872944), (('one', 'side'), 6.6622295748656075), (('be', 'seen'), 6.541074045203153), (('you', "'"), 6.516520810844119), (('stood', 'there'), 6.448805617251043), (('i', "'"), 6.419659271591531), ((':', '"'), 6.3780157632491115), (('"', '.'), 6.378015763249108), (('one', 'day'), 6.340301479978244), (('ll', 'be'), 6.316007489568383), (('lay', 'there'), 6.271267431698858), (('"', 'oh'), 6.25248488116525), (('next', 'room'), 6.20909228565187), (('her', 'face'), 6.097731176704167), (('at', 'home'), 6.048527178703397), (('more', 'than'), 6.023421479441424), (('did', 'not'), 6.0152690165121925), (('no', 'one'), 5.997096725460468), (('from', 'under'), 5.9911595125061226), (("gregor's", 'father'), 5.965389358070011), (('aware', 'that'), 5.910868451355775), (('would', 'often'), 5.910247803877549), (('much', 'more'), 5.858362233170929), (('be', 'heard'), 5.830580662398139), (("gregor's", 'mother'), 5.809961177282389), (('of', 'drawers'), 5.807475302135352), (('of', 'course'), 5.80747530213535), (('kind', 'of'), 5.807475302135348)]

## Death of a Salesman:

[(('ebbets', 'field'), 12.756243565712321), (('con', 'ﬁdence'), 12.171281064991167), (('terri', 'ﬁc'), 12.171281064991167), (('miss', 'forsythe'), 11.640766348292384), (('high', 'school'), 10.88177444779618), (('wall', 'line'), 10.586318564270009), (('slight', 'pause.'), 10.503262824542452), (('mrs.', 'loman'), 10.483225071305904), (('slight', 'pause'), 10.296811947075025), (('y', 'know.'), 10.09327855298989), (('new', 'england'), 10.019277971546117), (('bill', 'oliver'), 10.019277971546115), (('new', 'york'), 10.019277971546114), (('four', 'years'), 9.586318564270009), (('old', 'con'), 9.448815040520074), (('living', 'room.'), 9.441201856410197), (('any', 'more.'), 9.43792672437734), (('living', 'room'), 9.296811947075025), (('1', '0'), 9.285923630932286), (('twenty', 'thousand'), 9.245281646434943), (('business', 'world'), 9.171281064991165), (('mr.', 'loman'), 9.161296976418543), (('better', 'than'), 9.154679097654034), (('arm', 'around'), 9.089066301702978), (('an', 'order'), 8.881774447796179), (('will', 'ya'), 8.83341142623478), (('as', 'though'), 8.796885550209666), (('thousand', 'dollars'), 8.792769441737436), (('breaks', 'off'), 8.747254782485067), (('good', 'bye'), 8.690154375254549), (('calling', 'after'), 8.6772922243175), (('w', 'o'), 8.458563017071638), (('h', 'o'), 8.458563017071635), (('o', 'f'), 8.458563017071635), (('kitchen', 'table'), 8.434315470824957), (('ﬁrst', 'thing'), 8.36621795444155), (('years', 'old'), 8.349279366969158), (('o', 'clock'), 8.26591793912924), (('l', 'e'), 8.24317398347295), (('good', 'night.'), 8.13761335222577), (('two', 'hundred'), 8.100891737099767), (('talking', 'about'), 8.06847025940739), (('sitting', 'down'), 7.95894230831864), (('your', 'father'), 7.951533364910281), (('sit', 'down'), 7.8520271044021275), (('shut', 'up'), 7.716227886864441), (('this', 'morning'), 7.7118494463538685), (('how', 'much'), 7.697349876658752), (('looks', 'around'), 7.6561068944268715), (('well', 'liked'), 7.6547055392502585)]

Something particular about these two stories is that they both take place mostly in a single setting. Metamorphosis easily could have taken place on the stage as a play due to the minimum change of scenery from the inside of a single apartment. Metamorphosis has popular bigrams in both the regular and mutual information score distributions to include pairs like living room, next room, at home, and other side (referring to the door that separates Gregor from the other characters in the story). Similarly, Death of a Salesman has word pairs such as new england, new york, and kitchen table.

While these stories largely share high scores for bigrams having to do with scenery and settings, they also both heavily focus on family. For family, two of the most frequent bigrams focus on Gregor’s relationship to his mother and father by referring to them as gregor’s mother and gregor’s father instead of calling them by their names. Another common bigram for Metamorphosis is my parents. For Death of a Salesman, the characters are mostly referred to by their names. However, there’s at least one pair of highly associated words that point to Mrs. Lowman, the main characters wife, talking to his grown sons about him using the two words “your father”.

I’m sure many more similar word associations could be found between these two stories. However, at the very least the word frequencies and bigram scores seem to paint a pretty clear picture of how these families associate with one another through money, space, and personal familial relationships. AT the very least we can see a skeleton of the story these words and bigrams come from by reading through these lists.