Introduction

In this section, we applied the Named Entity Recognition technique in Neo4j and tried to get conclusions about what types of words(entities) to quote can make a speech great and what differences that females and males focus on are.

codes

* Import the csv file, create the nodes and relationships

LOAD CSV FROM 'file:///speeches2.csv' AS line FIELDTERMINATOR ';'

with line

skip 1

CREATE (s:Speech {Title:line[3], Content: line[6]})

CREATE (o:Orator{Title:line[3], Content: line[6], Gender: line[2]})

CREATE (t:Title {Title:line[3], Content: line[6]})

MERGE (s) -[:GIVEN\_BY]-> (o)

MERGE (s) -[:KNOWN\_AS]-> (t)

* Create an entity stream based on google’s NLP

MATCH (s:Speech)

CALL apoc.nlp.gcp.entities.stream(s, {

key: "AIzaSyDMu-yXIK5QrmEs3hjAdiTU80PdM2JZebM",

nodeProperty: "Content",

write:TRUE

})

YIELD value

UNWIND value.entities AS entity

RETURN entity;

* Connect entities to speeches

MATCH (s:Speech)

CALL apoc.nlp.gcp.entities.stream(s, {

key: "AIzaSyDMu-yXIK5QrmEs3hjAdiTU80PdM2JZebM",

nodeProperty: "Content"

})

YIELD value

UNWIND value.entities AS entity

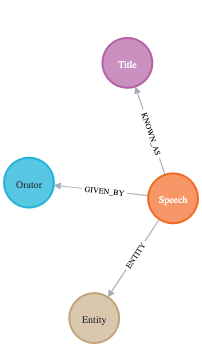
MERGE (e:Entity {name: entity.name})

SET e.type = entity.type

MERGE (s)-[:ENTITY]->(e);

* Check the visualisation of nodes and relationships

call db.schema.visualization()



* Check the frequency of different types of entities in total

MATCH (s:Speech)-[r:ENTITY]->(e:Entity)

with e.type as type, count(DISTINCT e) as number

return type, number

order by number desc

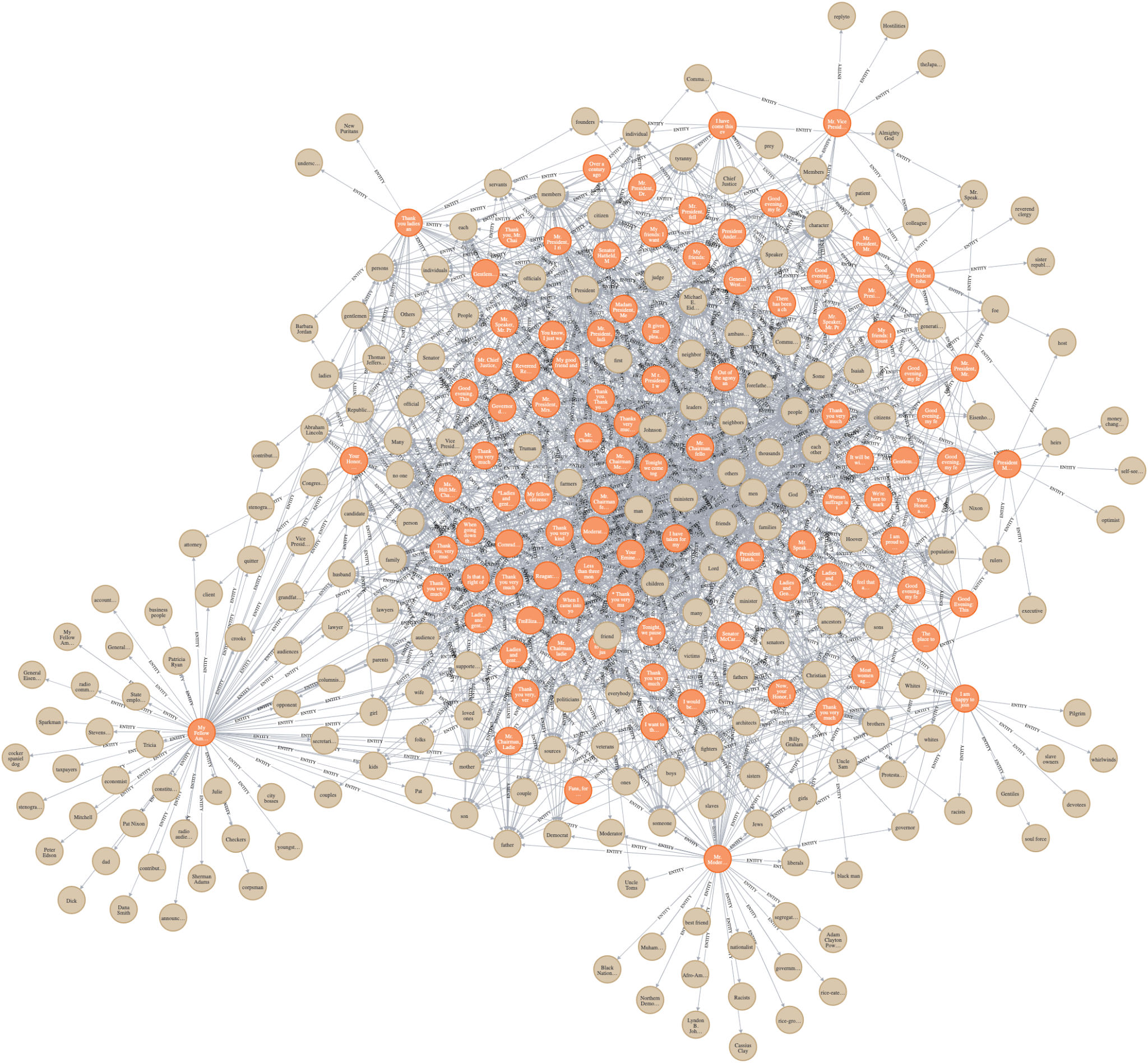
| **type** | **number** |
| --- | --- |
| **OTHER** | 6492 |
| **PERSON** | 2091 |
| **LOCATION** | 1134 |
| **ORGANISATION** | 704 |
| **NUMBER** | 554 |
| **EVENT** | 505 |
| **WORK\_OF\_ART** | 274 |
| **CONSUMER\_GOOD** | 146 |
| **PRICE** | 116 |
| **DATE** | 80 |
| **PHONE\_NUMBER** | 1 |
| **ADDRESS** | 1 |

* Visualise the the most used type of entity and its related words(except type: other)

MATCH (s:Speech)-[r:ENTITY]->(e:Entity)

where e.type='PERSON'

return s,r,e



* Check the frequency of different types of entities used by male

match (o:Orator)<-[r:GIVEN\_BY]-(s:Speech)-[t:ENTITY]->(e:Entity)

where o.Gender="Male"

with e.type as male\_type, count(DISTINCT e) as male\_number

return male\_type, male\_number

order by male\_number desc

| **male\_type** | **male\_number** |
| --- | --- |
| **OTHER** | 5810 |
| **PERSON** | 1725 |
| **LOCATION** | 1032 |
| **ORGANIZATION** | 597 |
| **NUMBER** | 495 |
| **EVENT** | 433 |
| **WORK\_OF\_ART** | 216 |
| **CONSUMER\_GOOD** | 126 |
| **PRICE** | 110 |
| **DATE** | 54 |
| **PHONE\_NUMBER** | 1 |
| **ADDRESS** | 1 |

* Check the frequency of different types of entities used by female

match (o:Orator)<-[r:GIVEN\_BY]-(s:Speech)-[t:ENTITY]->(e:Entity)

where o.Gender="Female"

with e.type as female\_type, count(DISTINCT e) as female\_number

return female\_type, female\_number

order by female\_number desc

| **female\_type** | **female\_number** |
| --- | --- |
| **OTHER** | 2064 |
| **PERSON** | 682 |
| **LOCATION** | 312 |
| **ORGANIZATION** | 209 |
| **EVENT** | 176 |
| **NUMBER** | 176 |
| **WORK\_OF\_ART** | 111 |
| **CONSUMER\_GOOD** | 34 |
| **DATE** | 26 |
| **PRICE** | 7 |

* Querying for entities (words) most used in total

MATCH (s:Speech)-[t:ENTITY]->(e:Entity)

WITH e.name as most\_used, count(e) as frequency

RETURN most\_used,frequency

ORDER by frequency desc

Table (top 50)

| **most\_used** | **frequency** |
| --- | --- |
| **people** | 95 |
| **one** | 95 |
| **all** | 89 |
| **men** | 85 |
| **life** | 85 |
| **country** | 82 |
| **world** | 82 |
| **man** | 79 |
| **way** | 77 |
| **nation** | 74 |
| **two** | 72 |
| **Book** | 71 |
| **CDs** | 70 |
| **Michael E. Eidenmuller** | 70 |
| **2008** | 70 |
| **McGraw-Hill** | 70 |
| **history** | 66 |
| **government** | 65 |
| **freedom** | 63 |
| **words** | 62 |
| **part** | 61 |
| **course** | 61 |
| **place** | 61 |
| **war** | 61 |
| **right** | 60 |
| **friends** | 60 |
| **God** | 60 |
| **power** | 60 |
| **things** | 60 |
| **others** | 59 |
| **children** | 58 |
| **peace** | 57 |
| **fact** | 57 |
| **society** | 57 |
| **nothing** | 56 |
| **lives** | 55 |
| **opportunity** | 55 |
| **some** | 55 |
| **women** | 54 |
| **work** | 54 |
| **order** | 54 |
| **purpose** | 52 |
| **home** | 52 |
| **President** | 51 |
| **three** | 51 |
| **many** | 50 |
| **hope** | 50 |
| **sense** | 50 |
| **nations** | 49 |
| **rights** | 48 |

* Querying for entities (words) most used by males

MATCH (o:Orator)<-[r:GIVEN\_BY]-(s:Speech)-[t:ENTITY]->(e:Entity)

WHERE o.Gender="Male"

WITH e.name as most\_usedbymen, count(e) as frequency

RETURN most\_usedbymen,frequency

ORDER by frequency desc limit 50

table(top50)

| **most\_usedbymen** | **frequency** |
| --- | --- |
| **people** | 80 |
| **one** | 78 |
| **all** | 75 |
| **life** | 71 |
| **country** | 70 |
| **men** | 68 |
| **world** | 67 |
| **man** | 66 |
| **nation** | 64 |
| **way** | 64 |
| **Book** | 58 |
| **Michael E. Eidenmuller** | 57 |
| **history** | 57 |
| **McGraw-Hill** | 57 |
| **two** | 57 |
| **2008** | 57 |
| **CDs** | 57 |
| **government** | 55 |
| **war** | 55 |
| **freedom** | 54 |
| **God** | 54 |
| **peace** | 53 |
| **part** | 52 |
| **power** | 52 |
| **right** | 51 |
| **words** | 51 |
| **course** | 51 |
| **place** | 50 |
| **fact** | 50 |
| **friends** | 49 |
| **lives** | 48 |
| **others** | 48 |
| **things** | 48 |
| **some** | 48 |
| **society** | 46 |
| **order** | 46 |
| **children** | 46 |
| **purpose** | 45 |
| **nothing** | 45 |
| **hope** | 44 |
| **three** | 43 |
| **sense** | 43 |
| **opportunity** | 43 |
| **nations** | 43 |
| **work** | 43 |
| **many** | 42 |
| **President** | 42 |
| **home** | 41 |
| **land** | 41 |
| **action** | 41 |

* Querying for entities (words) most used by females

MATCH (o:Orator)<-[r:GIVEN\_BY]-(s:Speech)-[t:ENTITY]->(e:Entity)

WHERE o.Gender="Female"

WITH e.name as most\_usedbywomen, count(e) as frequency

RETURN most\_usedbywomen,frequency

ORDER by frequency desc limit 50

Table (top 50)

| **most\_usedbywomen** | **frequency** |
| --- | --- |
| **one** | 17 |
| **men** | 17 |
| **women** | 16 |
| **world** | 15 |
| **people** | 15 |
| **two** | 15 |
| **all** | 14 |
| **life** | 14 |
| **CDs** | 13 |
| **2008** | 13 |
| **McGraw-Hill** | 13 |
| **man** | 13 |
| **way** | 13 |
| **woman** | 13 |
| **Book** | 13 |
| **Michael E. Eidenmuller** | 13 |
| **things** | 12 |
| **children** | 12 |
| **country** | 12 |
| **opportunity** | 12 |
| **home** | 11 |
| **work** | 11 |
| **family** | 11 |
| **words** | 11 |
| **place** | 11 |
| **society** | 11 |
| **nothing** | 11 |
| **others** | 11 |
| **friends** | 11 |
| **course** | 10 |
| **reason** | 10 |
| **law** | 10 |
| **thing** | 10 |
| **rights** | 10 |
| **nation** | 10 |
| **question** | 10 |
| **government** | 10 |
| **idea** | 9 |
| **times** | 9 |
| **history** | 9 |
| **right** | 9 |
| **something** | 9 |
| **struggle** | 9 |
| **mothers** | 9 |
| **President** | 9 |
| **freedom** | 9 |
| **faith** | 9 |
| **problems** | 9 |
| **education** | 9 |
| **part** | 9 |

Conclusions

1. Except for other entity type PERSON type is the most frequent entity type which appears in the speeches. Speakers also use location, organisation, number and event very frequently.
2. Compared to the frequent types of entity between females and males, most of the entity types have the same rankings. However, the entity type event appears to have a higher frequency ranking in female speeches than in male speeches.  
   Due to the entity types being broad and lacking context, we decided to further investigate by querying for frequency of words in all of the speeches, and then according to gender.
3. Person entity type like people, men, women, friends, children and president, location entity type like country, world, nation amd home, organisation entity type government and other type like history, life, society have the highest frequency in the great speeches.
4. Except for the common words It is highly likely that words like freedom, power, rights, god, peace, hope and opportunity can encourage the listeners more. We can also see that the word war appeared very often in the speeches which was also the main topic of the top 100 speeches in history.
5. From the word frequencies, we can indeed see that words like women, and home appear in the female speeches, but they do not appear in the male speeches. Children appear in both, but it is ranked lower in male word frequency table than in the female frequency table.

Notes

1. We are more interested in the capabilities of entity recognition and we tried to apply NER technique by python and spacy and then we found that it was much easier to do it in Neo4j because it was more straightforward to check the nodes and relationships and to visualise them.

Suggestions

1. Weaknesses; It is not very precise to derive broad generalisations from the frequency tables because a single article could throw off the categorisations.