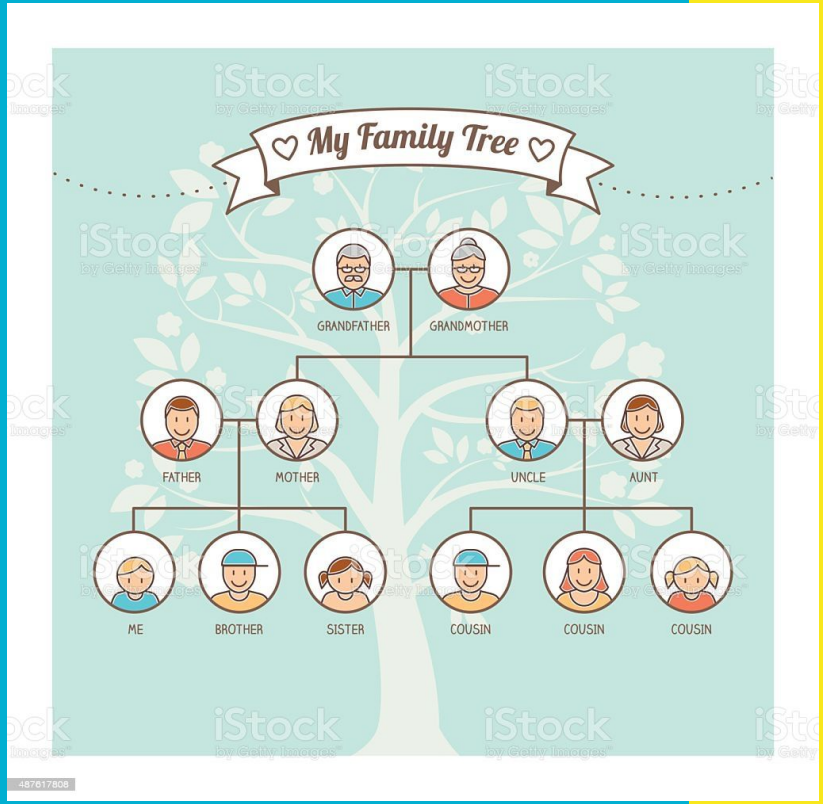
Geneology


My Family Tree
GRANDFATHER
GRANDMOTHER
FATHER
MOTHER
UNCLE
AUNT
ME
BROTHER
SISTER
COUSIN
COUSIN
COUSIN

Introduction

Project: Family Tree Extraction

1. Take in name
2. Lookup
3. Extract relevant sentences
4. Parse based on relation
5. Build data structure
6. Output GEDCOM, png

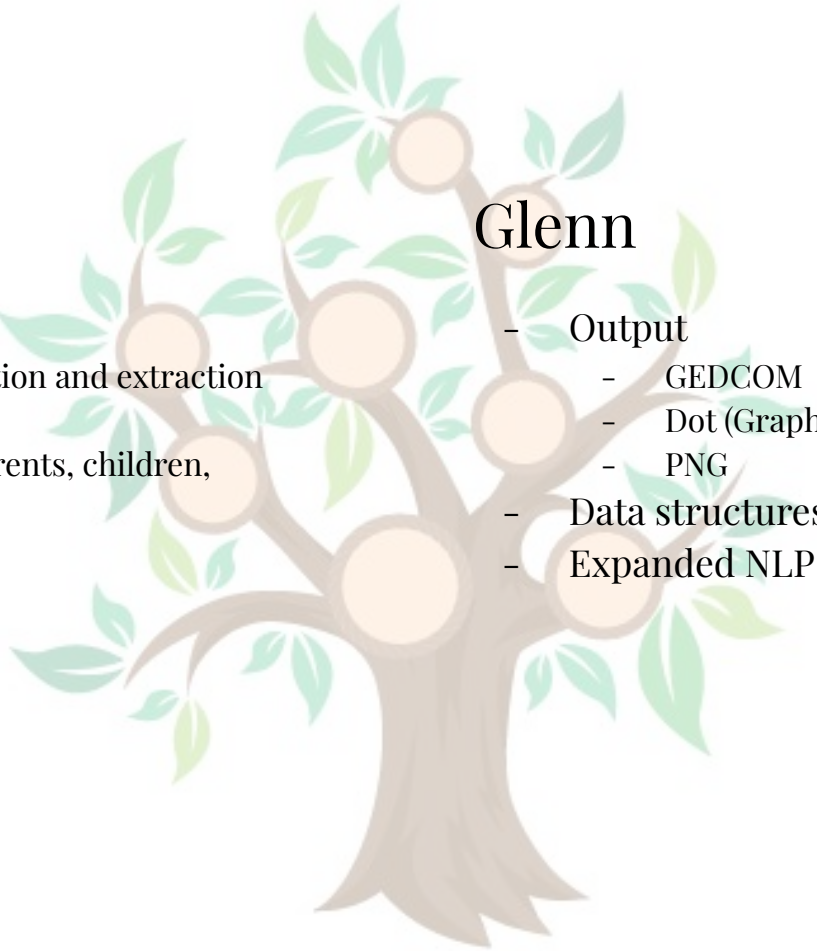
Distribution

Zach

- NLP
  - Sentence identification and extraction
  - Parsing
  - Building sets for parents, children, siblings

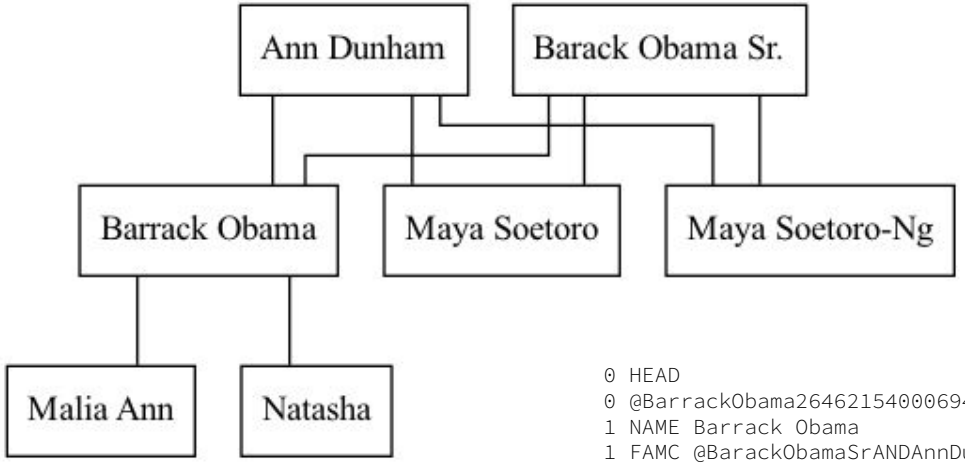
Glenn

- Output
  - GEDCOM
  - Dot (Graphviz)
  - PNG
- Data structures
- Expanded NLP work



Deliverable

- Input prompt
  - Person
- Output
  - output.ged
  - output.dot
  - output.png



output.png

```
0 HEAD
0 @BarrackObama264621540006948597021471759@ INDI
1 NAME Barrack Obama
1 FAMC @BarackObamaSrANDAnnDunham@
0 @BarackObamaSr2007700566239176808673841 31@ INDI
1 NAME Barack Obama Sr.
1 FAMS @BarackObamaSrANDAnnDunham@
0 @AnnDunham526942112921584512059546061 1946@ INDI
1 NAME Ann Dunham
1 FAMS @BarackObamaSrANDAnnDunham@
0 @BarackObamaSrANDAnnDunham@ FAM
1 HUSB @BarackObamaSr2007700566239176808673841 31@
1 WIFE @AnnDunham526942112921584512059546061 1946@
1 CHIL @BarrackObama264621540006948597021471759@
0 TRLR
```

output.ged

Results

Total: 77
Correct: 39
Not in family: 22
Wrong place: 1
Missed: 37

51% correct

Name	Total Family	Correct Placement	Not in family	Wrong Placement	Missed
Donald Trump	11	8	6	0	3
Barack Obama	5	5	0	0	0
Billy Ray Cyrus	12	4	3	0	8
Kim Kardashian	11	7	3	1	3
Kanye West	6	3	0	0	3
Steve Irwin	7	3	1	0	4
LeBron James	5	3	2	0	2
Will Smith	8	2	0	0	6
Adele	3	1	1	0	2
Joe Biden	9	3	6	0	6

Accomplishments

- Create family subtree
  - Able to output parents, siblings and children for given person
    - 2 error categories
      - Identification
      - Assignment
- Output to GEDCOM format
- Convert GEDCOM to graphical representation
- Minimal libraries used
  - Generalistic parsing and NLP techniques to capture a wide range of sentence structures and possible relation descriptions

Challenges

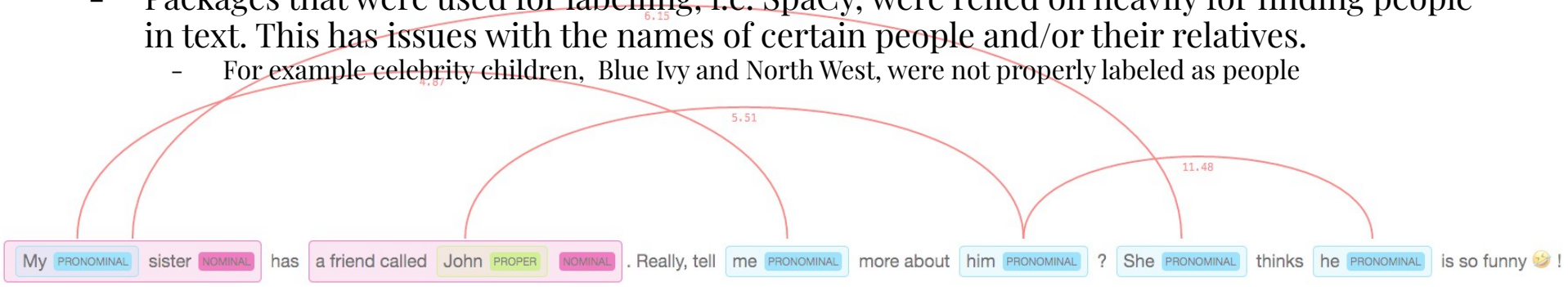
Several challenges presented themselves while working to solve this problem

Corpus

- Wikipedia was used as other data sources are less extensive or require subscriptions.
- No standard way for information to be entered in terms of formatting and word usage
  - Worked with assumption that He/Her/They pronouns refereed to person of interest when in some cases it referred to their parents or other people

Packages

- Hugging Face coreference resolution module could not be integrated into program due to interdependency issues and out of date modules used by Hugging faces
- Packages that were used for labelling, i.e. SpaCy, were relied on heavily for finding people in text. This has issues with the names of certain people and/or their relatives.
  - For example celebrity children, Blue Ivy and North West, were not properly labeled as people



Future work

Coreference resolution needs to be incorporated to improve accuracy and identify more relations

This will have to be done either through the creation of a new coreference module, or an updated version.

Continue fine tuning cases for sentence structure as found on wikipedia

Incorporate recursive searching to grow tree. Once the the accuracy of found trees are higher then the model could recurse on each member in the tree to rapidly expand it.

Demo