Lab_4

Loading Libraries

Part 1: Topic Modeling

```
[1] 6752    4
```

```
[1] "doc_id"      "screen_name" "party"       "message"
```

```
  doc_id     screen_name      party
1      1       EdJMarkey   Democrat
2      2 RepDrewFerguson Republican
3      3        RepJoshG   Democrat
4      4    RepWesterman Republican
5      5    lloyddoggett   Democrat
6      6      RepDelBene   Democrat
```

```
1
2
3                 Our economy needs a shot in the arm to spur growth.  As Co-Chair of the bipa
4
5 Even as families nationwide marveled at the science of the solar eclipse--
President Trump again rejected scientific counsel for his own pro-pollution agenda. His Admin
6                       The strength of our local economy depends on a well-maintained, modern
```

There are 6752 rows and 4 columns in this dataset. The name of the variables are "doc_id" "screen_name", "party", and "message".

2 (i) Corspus

2 (ii) Tokenization

2 (iii)

2 (iv)

```
Tokens consisting of 3 documents and 2 docvars.
1 :
 [1] "President"     "Trump"         "backs"         "Paris"
 [5] "Agreement"     "economic"      "environmental" "national"
 [9] "security"      "moral"         "disaster"      "United"
[ ... and 6 more ]

2 :
 [1] "Many"       "thanks"     "first"      "class"      "summer"
 [6] "interns"    "Washington" "hard"       "work"       "folks"
[11] "#GA03"

3 :
 [1] "economy"    "needs"      "shot"       "arm"        "spur"
 [6] "growth"     "Co-Chair"   "bipartisan" "Problem"    "Solvers"
[11] "Caucus"     "I've"
[ ... and 27 more ]
```

2 (v)

```
Document-feature matrix of: 6,752 documents, 21,029 features (99.86% sparse) and 2 docvars.
    features
docs president trump backs paris agreement economic environmental national
   1         1     1     1     1         1        1             1        1
   2         0     0     0     0         0        0             0        0
   3         0     0     0     0         0        0             0        0
   4         0     0     0     0         0        0             0        0
   5         1     2     0     0         0        0             0        0
   6         0     0     0     0         0        0             0        0
    features
docs security moral
   1        1     1
   2        0     0
   3        0     0
```

```
   4         0      0
   5         0      0
   6         0      0
[ reached max_ndoc ... 6,746 more documents, reached max_nfeat ... 21,019 more features ]
```

2 (vi)

```
[1] 5748 5484
```

After pre-processing, the resulting document-term matrix contains 5748 documents and 5484 unique terms. This means we retained 5748 Facebook posts that have at least 10 words and 5484 words that appear in at least 5 posts. This reduced set will be used for further analysis.

3 Topic Modeling

4

```
      Topic 1     Topic 2     Topic 3      Topic 4         Topic 5
 [1,] "tax"       "life"      "work"       "join"          "continue"
 [2,] "reform"    "history"   "together"   "town"          "proud"
 [3,] "plan"      "time"      "can"        "live"          "fight"
 [4,] "families"  "amendment" "come"       "hall"          "i'm"
 [5,] "cuts"      "story"     "hard"       "questions"     "support"
 [6,] "taxes"     "month"     "across"     "hope"          "keep"
 [7,] "code"      "right"     "way"        "tomorrow"      "fighting"
 [8,] "middle"    "state"     "done"       "please"        "protect"
 [9,] "class"     "read"      "working"    "social"        "stand"
[10,] "americans" "one"       "challenges" "hosting"       "allow"
[11,] "working"   "stories"   "good"       "event"         "also"
[12,] "bill"      "black"     "make"       "facebook"      "recent"
[13,] "gop"       "second"    "meet"       "constituents"  "member"
[14,] "money"     "freedom"   "nation"     "tune"          "i'll"
[15,] "pay"       "sense"     "video"      "tonight"       "everything"
      Topic 6     Topic 7        Topic 8        Topic 9        Topic 10
 [1,] "program"   "senator"      "community"    "america"      "states"
 [2,] "help"      "rights"       "city"         "country"      "united"
 [3,] "need"      "justice"      "texas"        "must"         "first"
 [4,] "services"  "court"        "university"   "nation"       "also"
 [5,] "support"   "senate"       "center"       "world"        "policy"
 [6,] "important" "judge"        "de"           "stand"        "always"
 [7,] "funding"   "law"          "san"          "daca"         "part"
```

```
 [8,] "provide"   "supreme"      "la"          "dreamers"    "many"
 [9,] "resources" "federal"      "el"          "around"      "nation's"
[10,] "grant"     "support"      "jewish"      "communities" "role"
[11,] "also"      "gorsuch"      "kansas"      "immigrants"  "glad"
[12,] "provides"  "constitution" "council"     "dream"       "team"
[13,] "critical"  "u.s"          "opportunity" "today"       "public"
[14,] "necessary" "record"       "building"    "us"          "throughout"
[15,] "including" "civil"        "california"  "status"      "strengthen"
      Topic 11     Topic 12     Topic 13      Topic 14      Topic 15
 [1,] "federal"    "county"     "american"    "law"         "small"
 [2,] "assistance" "community"  "people"      "department"  "business"
 [3,] "u.s"        "center"     "deserve"     "enforcement" "businesses"
 [4,] "puerto"     "local"      "better"      "safe"        "local"
 [5,] "rico"       "residents"  "congress"    "security"    "economy"
 [6,] "emergency"  "fire"       "put"         "police"      "communities"
 [7,] "hurricane"  "valley"     "process"     "communities" "farmers"
 [8,] "disaster"   "place"      "americans"   "immigration" "agriculture"
 [9,] "relief"     "area"       "made"        "officers"    "food"
[10,] "help"       "senior"     "clear"       "border"      "farm"
[11,] "efforts"    "officials"  "it's"        "local"       "owners"
[12,] "management" "evacuation" "coming"      "homeland"    "rural"
[13,] "recovery"   "north"      "whether"     "keep"        "help"
[14,] "fema"       "please"     "fact"        "line"        "across"
[15,] "support"    "post"       "republicans" "laws"        "growing"
      Topic 16     Topic 17          Topic 18      Topic 19    Topic 20
 [1,] "health"     "students"        "protect"     "need"      "honor"
 [2,] "care"       "school"          "free"        "crisis"    "service"
 [3,] "affordable" "education"       "information" "help"      "thank"
 [4,] "act"        "high"            "internet"    "opioid"    "day"
 [5,] "system"     "young"           "access"      "human"     "war"
 [6,] "insurance"  "congressional"   "can"         "must"      "serve"
 [7,] "access"     "student"         "open"        "drug"      "country"
 [8,] "obamacare"  "college"         "ensure"      "helping"   "honored"
 [9,] "healthcare" "congratulations" "protections" "epidemic"  "sacrifice"
[10,] "repeal"     "district"        "without"     "treatment" "award"
[11,] "improve"    "competition"     "rules"       "like"      "memorial"
[12,] "costs"      "career"          "control"     "abuse"     "serving"
[13,] "quality"    "schools"         "consumers"   "address"   "ceremony"
[14,] "premiums"   "opportunity"     "personal"    "21st"      "world"
[15,] "lower"      "app"             "online"      "combat"    "vietnam"
      Topic 21   Topic 22          Topic 23   Topic 24      Topic 25
 [1,] "national" "house"           "get"      "bill"        "women"
 [2,] "public"   "today"           "just"     "act"         "day"
```

```
 [3,] "park"     "vote"            "congress" "legislation" "today"
 [4,] "week"     "floor"           "can"      "house"       "men"
 [5,] "part"     "week"            "made"     "passed"      "every"
 [6,] "future"   "financial"       "long"     "bipartisan"  "across"
 [7,] "lands"    "spoke"           "time"     "senate"      "country"
 [8,] "natural"  "act"             "go"       "introduced"  "pay"
 [9,] "also"     "representatives" "past"     "pass"        "nation"
[10,] "native"   "republicans"     "issue"    "h.r"         "equal"
[11,] "alaska"   "voted"           "way"      "bills"       "paid"
[12,] "river"    "weeks"           "enough"   "system"      "uniform"
[13,] "concerns" "wall"            "now"      "process"     "fought"
[14,] "western"  "resolution"      "must"     "colleagues"  "equality"
[15,] "news"     "white"           "focus"    "support"     "planned"
      Topic 26        Topic 27    Topic 28      Topic 29    Topic 30
 [1,] "office"        "state"     "great"       "family"    "lives"
 [2,] "washington"    "one"       "discuss"     "great"     "violence"
 [3,] "district"      "many"      "yesterday"   "happy"     "lost"
 [4,] "staff"         "community" "issues"      "today"     "thoughts"
 [5,] "week"          "project"   "meeting"     "everyone"  "first"
 [6,] "visit"         "able"      "morning"     "thank"     "victims"
 [7,] "d.c"           "important" "talk"        "time"      "gun"
 [8,] "learn"         "army"      "association" "hope"      "prayers"
 [9,] "please"        "ryan"      "thank"       "team"      "families"
[10,] "information"   "paul"      "met"         "thanks"    "remember"
[11,] "dc"            "step"      "importance"  "friends"   "attack"
[12,] "capitol"       "speaker"   "thanks"      "birthday"  "americans"
[13,] "constituents"  "much"      "enjoyed"     "best"      "never"
[14,] "congressional" "corps"     "meet"        "wonderful" "steve"
[15,] "website"       "efforts"   "members"     "celebrate" "responders"
      Topic 31  Topic 32   Topic 33        Topic 34         Topic 35
 [1,] "it’s"    "military" "congress"      "president"      "us"
 [2,] "one"     "security" "congressman"   "trump"          "like"
 [3,] "just"    "north"    "rep"           "administration" "see"
 [4,] "right"   "u.s"      "read"          "donald"         "know"
 [5,] "going"   "defense"  "members"       "trump's"        "back"
 [6,] "now"     "korea"    "congressional" "trump’s"        "take"
 [7,] "said"    "national" "colleagues"    "executive"      "let"
 [8,] "think"   "strategy" "joined"        "order"          "time"
 [9,] "stop"    "world"    "letter"        "j"              "well"
[10,] "got"     "armed"    "bipartisan"    "obama"          "one"
[11,] "good"    "threat"   "caucus"        "white"          "many"
[12,] "we’re"   "region"   "john"          "actions"        "around"
[13,] "say"     "forces"   "friend"        "signed"         "weekend"
```

```
[14,] "time"    "must"     "democratic"    "ban"            "show"
[15,] "another" "south"    "full"          "action"         "put"
      Topic 36     Topic 37        Topic 38        Topic 39  Topic 40
 [1,] "u.s"        "committee"     "jobs"          "years"   "investigation"
 [2,] "service"    "hearing"       "economic"      "last"    "director"
 [3,] "air"        "house"         "job"           "year"    "general"
 [4,] "force"      "watch"         "create"        "night"   "russia"
 [5,] "academy"    "chairman"      "workers"       "since"   "russian"
 [6,] "military"   "today"         "economy"       "two"     "intelligence"
 [7,] "west"       "commerce"      "growth"        "one"     "sessions"
 [8,] "cancer"     "subcommittee"  "opportunities" "ago"     "independent"
 [9,] "learn"      "member"        "development"   "old"     "fbi"
[10,] "coast"      "including"     "trade"         "three"   "attorney"
[11,] "virginia"   "chamber"       "fair"          "week"    "special"
[12,] "guard"      "congresswoman" "grow"          "news"    "former"
[13,] "interested" "well"          "workforce"     "nearly"  "democracy"
[14,] "click"      "also"          "labor"         "another" "election"
[15,] "national"   "oversight"     "companies"     "example" "security"
      Topic 41     Topic 42      Topic 43     Topic 44         Topic 45
 [1,] "forward"    "americans"   "veterans"   "new"            "families"
 [2,] "look"       "health"      "va"         "funding"        "children"
 [3,] "working"    "republicans" "benefits"   "infrastructure" "home"
 [4,] "important"  "millions"    "care"       "programs"       "every"
 [5,] "work"       "care"        "receive"    "budget"         "opportunity"
 [6,] "secretary"  "bill"        "ensure"     "million"        "give"
 [7,] "looking"    "coverage"    "medical"    "critical"       "working"
 [8,] "leadership" "million"     "department" "billion"        "many"
 [9,] "check"      "senate"      "veteran"    "appropriations" "raise"
[10,] "move"       "insurance"   "affairs"    "including"      "better"
[11,] "dr"         "people"      "provide"    "state"          "kids"
[12,] "king"       "healthcare"  "access"     "mexico"         "life"
[13,] "life"       "medicaid"    "help"       "transportation" "child"
[14,] "last"       "republican"  "get"        "fund"           "now"
[15,] "legacy"     "conditions"  "quality"    "increase"       "support"
      Topic 46    Topic 47     Topic 48        Topic 49         Topic 50
 [1,] "today"     "make"       "energy"        "federal"        "can"
 [2,] "water"     "can"        "future"        "government"     "see"
 [3,] "take"      "need"       "change"        "congress"       "find"
 [4,] "ensure"    "sure"       "climate"       "regulations"    "open"
 [5,] "keep"      "want"       "industry"      "spending"       "plan"
 [6,] "safety"    "like"       "progress"      "use"            "new"
 [7,] "best"      "making"     "power"         "dollars"        "get"
 [8,] "rule"      "better"     "environment"   "agencies"       "visit"
```

```
 [9,] "michigan"  "still"      "clean"         "debt"           "sign"
[10,] "safe"      "heard"      "science"       "sexual"         "now"
[11,] "lake"      "island"     "space"         "accountability" "family"
[12,] "steps"     "calls"      "technology"    "assault"        "help"
[13,] "prevent"   "concerned"  "real"          "taxpayer"       "today"
[14,] "important" "voices"     "epa"           "regulatory"     "enrollment"
[15,] "clean"     "difference" "environmental" "budget"         "december"
```

```
      Topic 1     Topic 9       Topic 40        Topic 45      Topic 48
 [1,] "tax"       "america"     "investigation" "families"    "energy"
 [2,] "reform"    "country"     "director"      "children"    "future"
 [3,] "plan"      "must"        "general"       "home"        "change"
 [4,] "families"  "nation"      "russia"        "every"       "climate"
 [5,] "cuts"      "world"       "russian"       "opportunity" "industry"
 [6,] "taxes"     "stand"       "intelligence"  "give"        "progress"
 [7,] "code"      "daca"        "sessions"      "working"     "power"
 [8,] "middle"    "dreamers"    "independent"   "many"        "environment"
 [9,] "class"     "around"      "fbi"           "raise"       "clean"
[10,] "americans" "communities" "attorney"      "better"      "science"
[11,] "working"   "immigrants"  "special"       "kids"        "space"
[12,] "bill"      "dream"       "former"        "life"        "technology"
[13,] "gop"       "today"       "democracy"     "child"       "real"
[14,] "money"     "us"          "election"      "now"         "epa"
[15,] "pay"       "status"      "security"      "support"     "environmental"
```

Topic 1: Tax Reformation

Topic 1 is about tax reformation. The main words in this topic are "tax," "reform," "plan," "families," and "middle class." I chose this topic because tax reform is an important issue that affects families, especially the middle class, and it's often debated in American Congress (also in western countries). Additionally, I noticed a lot of content about U.S. taxes and taxpayers' money on social media, which made this topic interesting to explore.

Topic 9: Political Rhetoric

Topic 9 is about political rhetoric. It includes words like "America," "nation," "stand," "dreamers," and "immigrants." I chose this topic because immigration has been a key political issue, especially related to policies about Dreamers and the role of immigrants in the country. Considering the upcoming U.S. election, American politicians often address these issues in their speeches or interviews using phrases like "we are the greatest nation in the world" or "where do we stand today compared to our great history," which makes this topic particularly relevant.

Topic 40: Russian Invlovement

It includes words like “investigation,” “Russia,” “FBI,” and “intelligence.” This was a major political issue in 2017 following the U.S. election, regarding the alleged Russian influence on the election. I chose this topic because of its significance to U.S. politics and society.

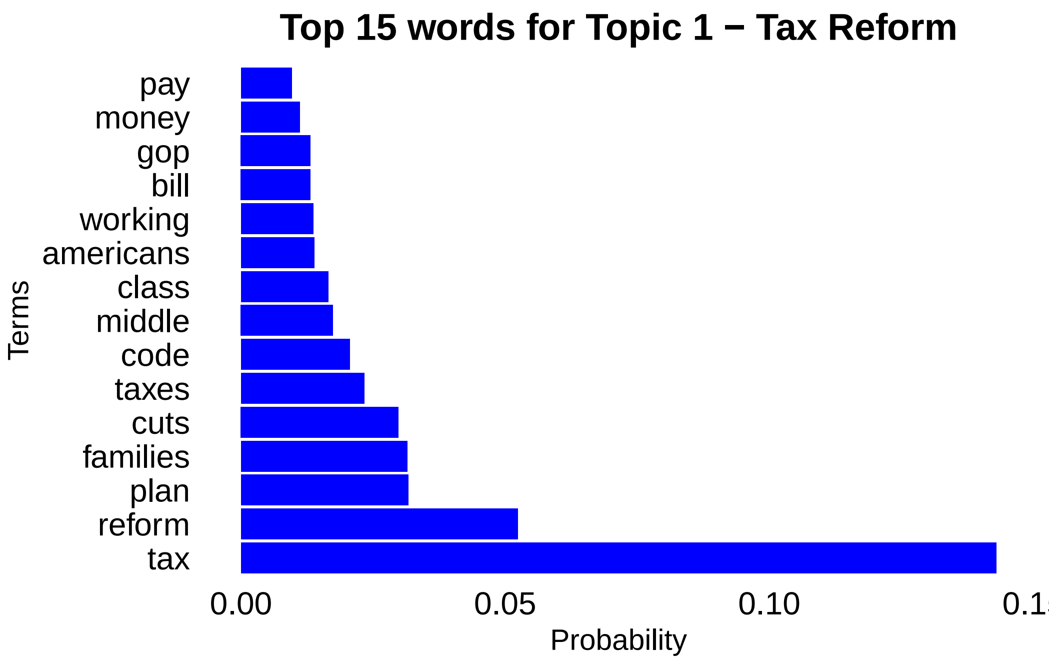
Topic 45: Social Welfare

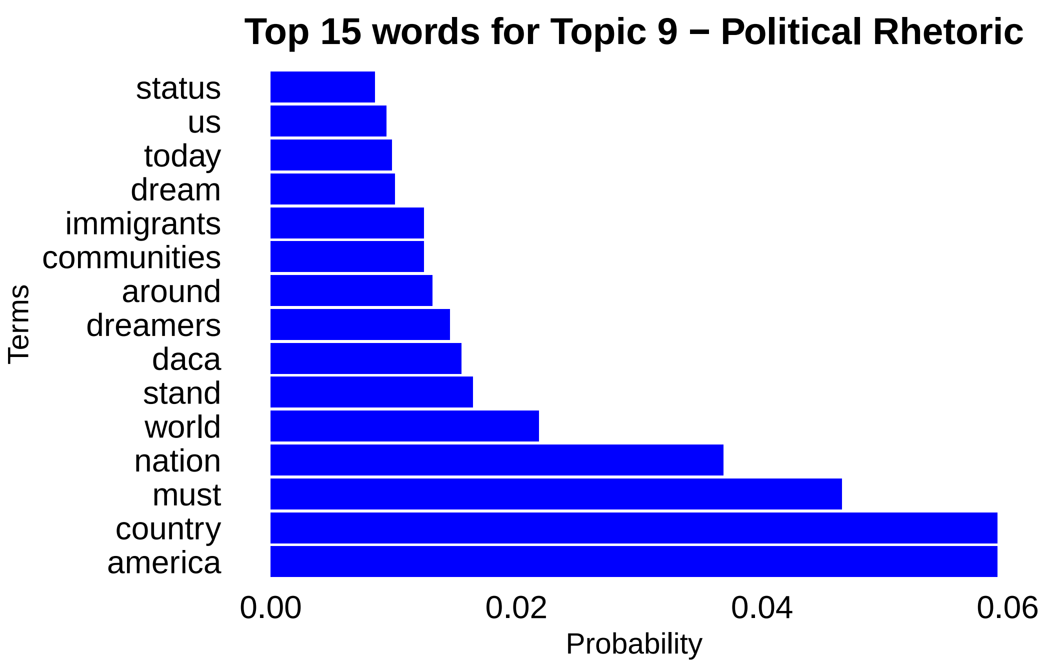
The main words here are “families,” “children,” “home,” “opportunity,” and “working.” I chose this topic because the well-being of families and children is a key focus of social welfare policies. It is important to explore policies that provide support for family life, improve living conditions, and create opportunities for children to thrive, especially in relation to social welfare programs aimed at helping working families.

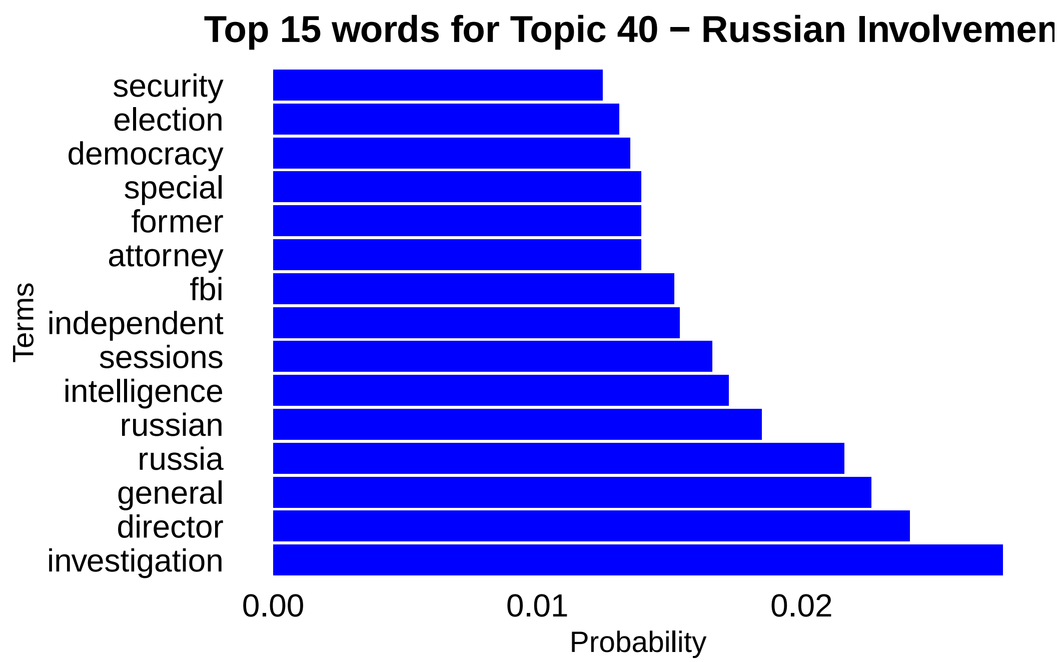
Topic 48: Sustainable Development

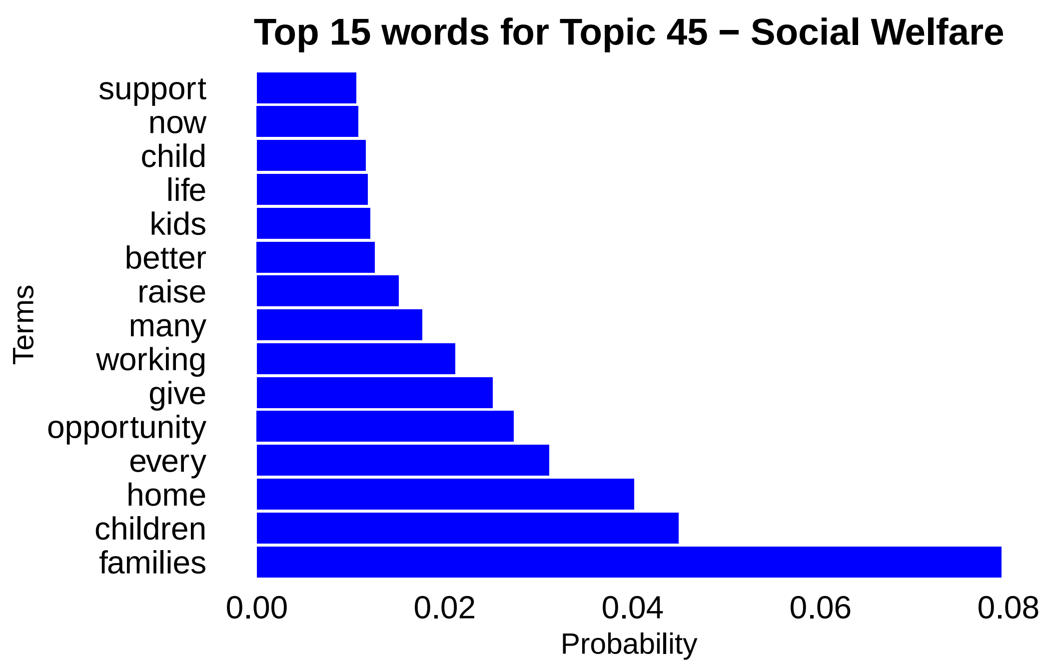
The main words in this topic include “energy,” “future,” “climate,” “industry,” “progress,” “power,” “environment,” and “clean.” These terms suggest a focus on sustainability, environmental protection, and clean energy. The inclusion of words like “science,” “technology,” and “environmental” highlights the role of innovation in addressing climate change and advancing clean energy solutions. I labeled this topic Sustainable development because it reflects discussions around environmental progress, climate change, and the future of clean energy industries.

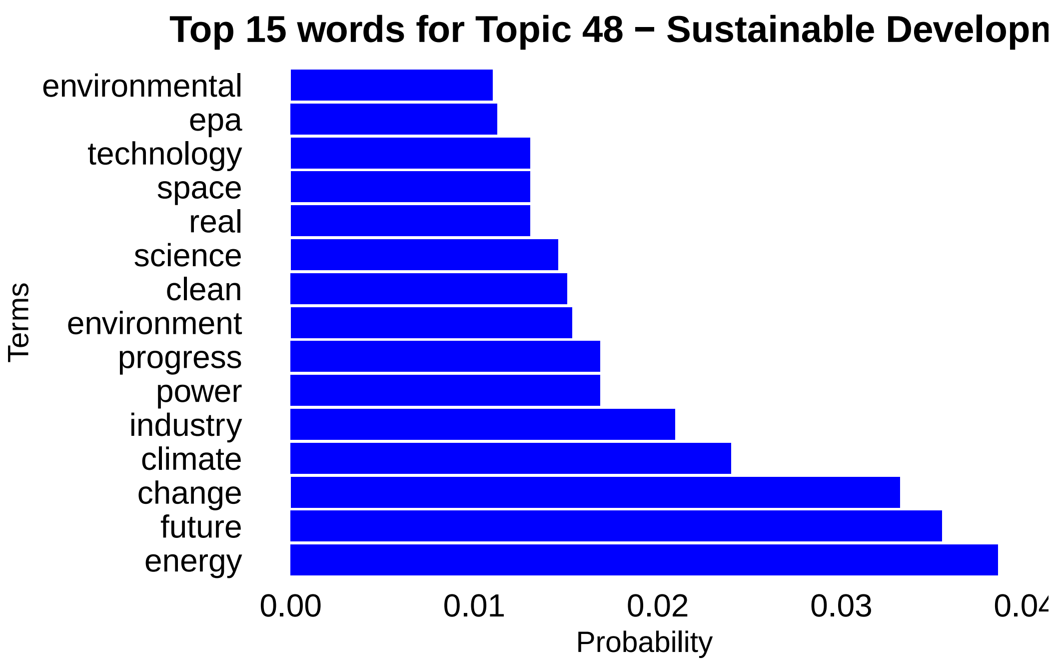
Bar Charts for Each Topic






Top 15 words for Topic 40 − Russian Involvemen
Terms
security
election
democracy
special
former
attorney
fbi
independent
sessions
intelligence
russian
russia
general
director
investigation
0.00
0.01
0.02
Probability


Top 15 words for Topic 45 − Social Welfare
Terms
support
now
child
life
kids
better
raise
many
working
give
opportunity
every
home
children
families
0.00
0.02
0.04
0.06
0.08
Probability



General Assesment

In my opinion, among all the topics, there are a few "junk" topics, such as topic 47, topic 26, topic 21, topic 23, topic 35, and topic 8. In these topics, the words seem random and do not fit together in a meaningful way. For example, in topic 8, the words include "Jewish," "Texas," "university," "de," and "San," which don't form a clear theme. Similarly, in topic 57, most of the words are verbs like "make," "can," "making," and "calling," which don't add much to a specific topic.

However, most of the topics make sense and are clearly themed. In my selected topics, the words are distinct and meaningful. For instance, in the topic I labeled "Sustainable Development" , words like "environment," "climate," "energy," "industry," and "future" all relate well to sustainability and environmental discussions. In the "Tax Reform" topic, words like "pay," "money," "tax," "families," and "working" are all clearly connected to tax-related issues.

The "Political Rhetoric" topic has some relevant words like "country," "world," "nation," and "immigrants," though I was hoping for more distinct terms. Still, it captures the idea of political discourse in the U.S. In the "Russian Involvement" topic, the words are clearly focused, with terms like "democracy," "Russian," "intelligence," "security," and "investigation," all reflecting the discussions around alleged Russian influence in the 2017 U.S. election.

Finally, the "Soacial Welfare" topic is also well-themed, with words like "child," "kids," "support," "better," and "raise," all related to family and social well-being.

5

```
[1] "Top 3 documents for Russian Investigation (Topic 40):"
```

```
[1] 3722 2084 1519
```

```
[1] "Top 3 documents for Sustainable Development (Topic 48):"
```

```
[1]  497 4290 1394
```

```
[1] "Top 3 texts for Russian Influence:"
```

```
[1] "For more than 6 years, Speaker Paul Ryan and Congressional Republicans have said one thi
[2] "Allowing gun owners to purchase suppressors absent federal regulation is important not j
[3] "In a few minutes, I will be on News Talk 1230, The Talk of Waco‘s the James Show to give
```

```
[1] "Top 3 texts for Sustainable Development:"
```

```
[1] "After I introduced legislation to reform the VA and ensure bad employees can be held acc
[2] "Guess who sent us the biggest, hugest, beautifulest ever ever Christmas card-
ever, ever?   You guessed it-it’s from the only one who has the really really biggest best mo
[3] "Wonderful morning yesterday spent visiting Lutheran Services Florida's Head Start center
making sure they all have an opportunity to succeed! Proud to support their continued work in
```

Based on the texts retrieved for topic 40 and topic 48, the documents do not seem to match the expected themes. For topic 40, the documents talk about healthcare, tax policies, firearm suppressors, and a general Washington update, which don’t align with the idea of the Russian involvement. This suggests that the topic model may have misclassified these documents or that they are only loosely connected to the theme. For topic 48, the content is mixed. One document talks about reforms to the Veterans Affairs system, which is not directly related to sustainability, and another is a lighthearted note about receiving a large Christmas card. However, the third document discusses supporting families and building a strong future for children, which somewhat aligns with social development, though it focuses more on social welfare than environmental sustainability. Overall, while the top words for these topics are clear, the documents retrieved do not always reflect those themes accurately, suggesting the topic model might not have captured the core ideas well.

6. Topic Model with K = 3

```
      Topic 1     Topic 2         Topic 3
 [1,] "president" "veterans"      "health"
 [2,] "today"     "great"         "care"
 [3,] "trump"     "community"     "bill"
 [4,] "day"       "week"          "act"
 [5,] "country"   "u.s"           "tax"
 [6,] "us"        "office"        "americans"
 [7,] "must"      "today"         "can"
 [8,] "congress"  "thank"         "families"
 [9,] "years"     "service"       "work"
[10,] "law"       "district"      "american"
[11,] "first"     "washington"    "need"
[12,] "time"      "congressional" "house"
[13,] "women"     "see"           "new"
[14,] "security"  "local"         "make"
[15,] "states"    "students"      "people"
```

"Topic 1" can be labeled as "Politics and Governance". The words include "president," "trump," "congress," "country," and "law," which all relate to political discussions and national governance. The words make sense together, and the topic is clearly focused on politics and security.

"Topic 2" can be labeled as "Community and Veterans". The main words are "veterans," "community," "service," and "district." These words suggest discussions about veterans' services and community-related topics. While it's a bit broad, the words fit well together. I believe this is related to war veterans and involves engaging them in various community services or initiating several community programs to make their lives more convenient when they return from overseas.

"Topic 3" can be labeled as "Healthcare and Tax Reform". The words "health," "care," "bill," "tax," and "families" point to discussions about healthcare policies and tax reform. These words might make sense together as healthcare and taxes are common issues debated in relation to families and government policy.

Overall, I think the topics in the K = 3 model are broad but still make sense. Each one has a clear focus: politics, community services, and healthcare/tax reform. However, they are more general compared to the K = 50 model, which captured more specific themes.

Which of the two K's do I prefer?

The K = 50 model captures specific themes with more detail, as seen in the top words for topics like Russian Investigation and Sustainable Development. However, when we retrieved the top documents for these topics, the documents did not match the expected themes. This suggests that while the model did a good job defining topics, it struggled with correctly classifying the documents. On the other hand, the K = 3 model is simpler and easier to interpret, with broader themes that cover larger topics. It did not provide the same level of detail as the K = 50 model, but it was more straightforward in terms of topic labeling. The downside of K = 3 is that it mixes different subjects into the same topic, making it less useful for capturing specific discussions. Considering these outcomes of our analysis, I believe it would be a good idea to try more values for K to see which one works best in terms of both defining clear themes and correctly classifying the documents. This approach might help find a balance between detail and accuracy in document classification.

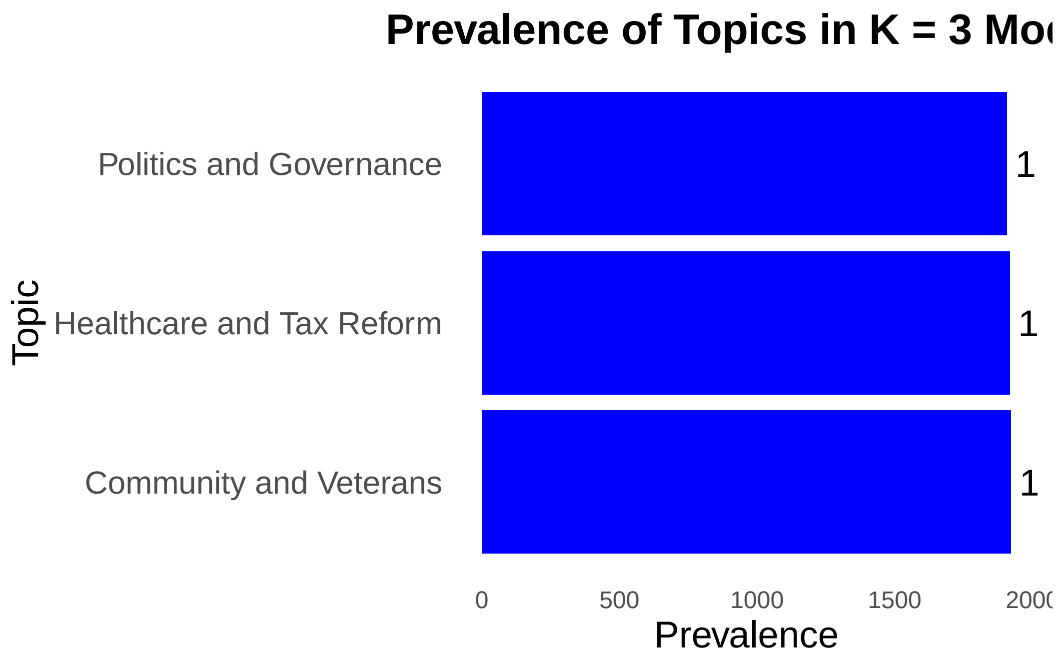
But between these two K values, I would prefer K = 3 because the words are more distinct and clear compared to K = 50.

7 (i)

```
[1] "The most prevalent topic is Topic 2"
```

```
       1        2        3
1907.780 1922.435 1917.785
```

Creating a Bar chart



The bar chart shows how frequently each of the three topics appears across all the documents in the K = 3 model. The three topics are Politics and Governance, Healthcare and Tax Reform, and Community and Veterans. The prevalence, or frequency, of each topic is shown on the x-axis, and all three topics have similar values, close to 1900. This means that each of these topics is discussed about equally in the dataset, with no one topic standing out as being much more important than the others. The model seems to capture a balanced view of the discussions, dividing the content fairly evenly among these three topics.

7 (ii)

```
[1] "T-test for Politics and Governance (Topic 1)"


    Welch Two Sample t-test

data:  df_democrats$X1 and df_republicans$X1
t = 5.227, df = 5731.5, p-value = 1.784e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.00659578 0.01451231
```

```
sample estimates:
mean of x mean of y
0.3370646 0.3265105

[1] "T-test for Healthcare and Tax Reform (Topic 2)"


    Welch Two Sample t-test

data:  df_democrats$X2 and df_republicans$X2
t = -13.834, df = 5715.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03433549 -0.02581226
sample estimates:
mean of x mean of y
0.3197456 0.3498195

[1] "T-test for Community and Veterans (Topic 3)"


    Welch Two Sample t-test

data:  df_democrats$X3 and df_republicans$X3
t = 8.9242, df = 5730.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01523193 0.02380772
sample estimates:
mean of x mean of y
0.3431898 0.3236700
```

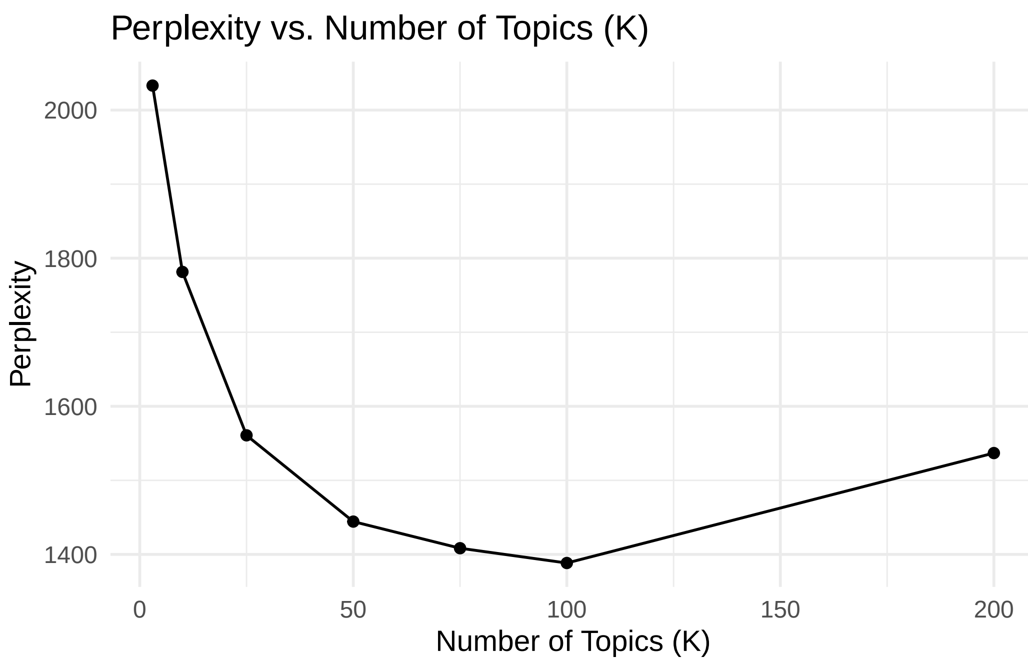
The results show significant differences in how Democrats and Republicans discuss certain topics. For Politics and Governance, Democrats have an average prevalence of 0.337, while Republicans have 0.327. Although Democrats talk about this topic slightly more than Republicans, the difference is small but statistically significant (p-value = 1.784e-07). For Healthcare and Tax Reform, Republicans discuss this topic much more frequently than Democrats, with an average prevalence of 0.350 for Republicans and 0.320 for Democrats. This difference is highly significant (p-value $<$ 2.2e-16). Lastly, for Community and Veterans, Democrats have an average prevalence of 0.343, while Republicans have 0.324. The difference here is also significant (p-value $<$ 2.2e-16). These results suggest that Republicans tend to focus more on healthcare and tax reform, while Democrats are more focused on community and veterans' issues.

Bonus Attempt

for loop

```
[1] "Computed perplexity for K = 3"
[1] "Computed perplexity for K = 10"
[1] "Computed perplexity for K = 25"
[1] "Computed perplexity for K = 50"
[1] "Computed perplexity for K = 75"
[1] "Computed perplexity for K = 100"
[1] "Computed perplexity for K = 200"
```

```
[1] 2033.089 1781.424 1560.784 1444.306 1408.347 1388.364 1536.865
```



Based on the plot of Perplexity vs. Number of Topics (K), we can interpret it as follows. As the number of topics increases from 3 to around 100, the perplexity decreases, showing that adding more topics helps the model better capture the structure of the data. Lower perplexity means better performance. The lowest perplexity is around K = 100, which indicates that 100 topics provide a good balance between having enough topics to capture details in the data without overfitting. After K = 100, perplexity starts to increase again, which means adding more topics, like K = 200, leads to overfitting or unnecessary complexity. In this case, the model may capture noise rather than useful patterns. So, K = 100 seems to be the most reasonable choice because it offers the best balance between model complexity and performance. Choosing a lower or higher K, such as 3 or 200, might result a less optimal model, as shown by the higher perplexity values.

Part 2: Word Embedding

1. Data Processing

2. Word Embeddings

```
   user  system elapsed 
 14.445   0.063  14.621 
```

3.

```
    today    health president      care     house       can      bill    people 
     1271      1068      1019       997       908       898       892       842 
      act  american       tax      work       new     trump  congress americans 
      832       828       797       782       712       698       695       693 
 veterans     great  families  national 
      688       688       665       634 
```

```
$president
       term1            term2 similarity rank
1  president   administration  0.8923684    1
2  president           donald  0.7739251    2
3  president            elect  0.7555957    3
4  president administration's  0.7398769    4
5  president           barack  0.7333671    5
6  president            admin  0.7253823    6
7  president      billionaire  0.7121655    7
8  president           pardon  0.7010574    8
9  president      impeachment  0.6989445    9
10 president          declare  0.6879105   10

$trump
   term1       term2 similarity rank
1  trump     trump's  0.8824441    1
2  trump       obama  0.8355765    2
3  trump     obama's  0.7978344    3
4  trump       elect  0.7395875    4
5  trump resignation  0.7282059    5
```

```
6  trump     actions  0.7278390    6
7  trump  outrageous  0.7197942    7
8  trump           j  0.7105138    8
9  trump        vice  0.7093837    9
10 trump    incoming  0.7006133   10

$american
      term1      term2 similarity rank
1  american  america's  0.7034523    1
2  american    america  0.6694897    2
3  american   nation's  0.6649768    3
4  american peacefully  0.6537287    4
5  american    profits  0.6408259    5
6  american   shameful  0.6248575    6
7  american      value  0.6118039    7
8  american      leads  0.6082526    8
9  american    country  0.6078990    9
10 american     houses  0.6073859   10
```

The results make sense overall. For the word "president," the nearest terms include "administration," "donald," "barack," and "impeachment," which are all related to the context of political leadership and recent U.S. presidents. This shows that the word embedding model has captured the relationships between these terms. For "trump," the closest words include "trump's," "obama," "elect," and "resignation," which are also relevant to political discussions, especially regarding Trump's presidency and related actions. Lastly, for "american," the nearest words include "freedom," "america's," "deserve," and "america," which all fit well with the theme of American identity and values, or the way it is written and discussed in the media or world. Overall, the results seem reasonable because the words identified as being similar are closely related to the focal terms in the context of U.S. politics.

4 (i)

```
[1] 5512   50
```

4 (ii)

4 (iii)

```
   term1         term2 similarity rank
1      A          king  1.0139786    1
2      A         woman  0.9219112    2
```

```
3       A         luther  0.8671174    3
4       A       suffrage  0.8547995    4
5       A           luke  0.8493186    5
6       A     basketball  0.8337910    6
7       A     nationally  0.8247301    7
8       A   demonstrates  0.8075623    8
9       A         legacy  0.7894152    9
10      A          harry  0.7866105   10
11      A           ncaa  0.7825570   11
12      A            hot  0.7819660   12
13      A              o  0.7810785   13
14      A     convention  0.7810330   14
15      A            day  0.7764297   15
16      A          catch  0.7624338   16
17      A         soccer  0.7587073   17
18      A          among  0.7575252   18
19      A         degree  0.7557658   19
20      A          frank  0.7546312   20
```

The result does not show "queen" in the top 20 most similar words to the kingtowoman vector. The words that appear, such as "luther," "suffrage," and "basketball," seem unrelated to the expected analogy. There are a few possible reasons for this. The dataset we are using, which consists of U.S. Congress Facebook posts, may not contain enough references to words like "king," "queen," "man," and "woman" in contexts that would allow the model to learn these classical analogy relationships. The word embeddings are trained on the available data, so if these relationships were not common in the dataset, the model might struggle to capture them. Words like "A", "O" or "laid" appearing in the top results could suggest that there is some noise in the embeddings, with frequently occurring words dominating the results. In this case, the model likely learned different relationships that are more relevant to U.S. politics and social discussions, rather than the kind of gender-role analogy represented by "king" and "queen." If we were working with a dataset more focused on literature or general knowledge, probably then we would be more likely to see "queen" appear in the analogy task.

4 (iv)

```
[1] "Number of embedding dimensions: 200"
```

```
[1] "Number of words with embedding vectors: 400000"
```

The pre-trained embedding model contains 200 embedding dimensions, meaning each word is represented by a 200-dimensional vector. Additionally, the model has embedding vectors for

400,000 words. This means that the model was trained on a large corpus and can represent 400,000 unique words with 200-dimensional vectors.

4 (v)

```
   term1     term2 similarity rank
1      A      king  0.4630444    1
2      A     queen  0.4272459    2
3      A  princess  0.4001166    3
4      A    prince  0.3994429    4
5      A    throne  0.3877838    5
6      A   emperor  0.3765438    6
7      A     royal  0.3732692    7
8      A  daughter  0.3721567    8
9      A   monarch  0.3719335    9
10     A   kingdom  0.3668355   10
11     A     crown  0.3652451   11
12     A    mother  0.3617875   12
13     A       her  0.3608811   13
14     A  marriage  0.3567054   14
15     A      wife  0.3565992   15
16     A elizabeth  0.3557918   16
17     A     woman  0.3553641   17
18     A   duchess  0.3549088   18
19     A adulyadej  0.3543496   19
20     A      duke  0.3536174   20
```

In the results from the pre-trained embedding model, "queen" appears as the second most similar word to the kingtowoman vector. This shows that the pre-trained model successfully captured the analogy where "king" is to "man" as "queen" is to "woman." Other related words like "princess," "prince," "throne," and "emperor" also appear in the top 20, which further shows the model understands relationships within royal and gender-related contexts. The pre-trained model performs better because it was trained on a large and diverse set of data from Wikipedia and news articles, which includes many examples of these words used together. In contrast, the self-trained model did not capture this relationship, likely because the dataset it was trained on (U.S. Congress Facebook posts) did not contain enough relevant examples. The pre-trained model has broader knowledge, which helps it understand these kinds of analogies better.

4 (vi)

Based on the focus of the Facebook/Congress model, which was trained on U.S. Congress Facebook posts, it is less likely to accurately capture occupational gender bias. This is because that dataset may not include enough varied occupations or gender-related discussions. On the other hand, the pre-trained GloVe model, which was trained on a large and diverse dataset like Wikipedia and news articles, would be much better at capturing these biases. The pre-trained model probably seen more examples of different occupations and gender roles, making it more reliable for calculating occupational gender bias. So, if we were to calculate the bias scores, the pre-trained model would likely give more accurate and meaningful results.

5

Process the data for democrats

For Republicans

Comparing Nearest Terms for Specific Words

```
$healthcare
        term1          term2 similarity rank
1  healthcare           care  0.7989002    1
2  healthcare         health  0.7894422    2
3  healthcare       coverage  0.7415851    3
4  healthcare            ACA  0.7397463    4
5  healthcare            out  0.6855110    5
6  healthcare      dismantle  0.6812945    6
7  healthcare    replacement  0.6807656    7
8  healthcare          scrap  0.6726055    8
9  healthcare PayMoreForLess  0.6713513    9
10 healthcare         choice  0.6670124   10

$healthcare
        term1       term2 similarity rank
1  healthcare     failing  0.7691137    1
2  healthcare      broken  0.7690855    2
3  healthcare        mean  0.7209541    3
4  healthcare replacement  0.7151047    4
5  healthcare        care  0.7097121    5
6  healthcare     reforms  0.6877396    6
7  healthcare      lowers  0.6875343    7
8  healthcare     replace  0.6822289    8
```

```
9  healthcare        AHCA  0.6749293    9
10 healthcare   repealing  0.6734539   10
```

I can see some differences, and they moderately align with my expectations. Democrats tend to use softer words like "coverage," "insurance," "health," "care," and "procedures" when discussing healthcare. These words reflect a focus on providing or improving access to healthcare, which aligns with how Democrats often present themselves.

On the other hand, Republicans tend to use more critical and stronger words when discussing healthcare. Words like "broken," "failing," "reforms," and "repealing" are more confrontational and align with the Republican stance on criticizing existing healthcare policies, especially the Affordable Care Act (ACA). This reflects their political rhetoric, which often emphasizes the need for change or dismantling of current systems.

```
$veteran
     term1       term2 similarity rank
1  veteran        vets  0.7873110    1
2  veteran     suicide  0.7822053    2
3  veteran   seriously  0.6980242    3
4  veteran        went  0.6867132    4
5  veteran       Medal  0.6829231    5
6  veteran      Flight  0.6606839    6
7  veteran  transition  0.6578178    7
8  veteran recognition  0.6540921    8
9  veteran     firearm  0.6518736    9
10 veteran        Hook  0.6432867   10
```

```
$veteran
     term1         term2 similarity rank
1  veteran          WWII  0.7353548    1
2  veteran        Flight  0.7069181    2
3  veteran           dog  0.6803447    3
4  veteran transitioning  0.6728607    4
5  veteran           Job  0.6713486    5
6  veteran    permission  0.6650827    6
7  veteran      Veterans  0.6649178    7
8  veteran          Bush  0.6589131    8
9  veteran         Small  0.6524770    9
10 veteran   appreciated  0.6488457   10
```

I can see some differences, and they moderately align with my expectations. Democrats tend to use words like "vets," "suicide," "veterans," "mission," and "wounds" when discussing veterans. These words suggest a focus on the well-being and support of veterans, addressing mental health and physical challenges they may face. On the other hand, Republicans use words like "WWII," "Flight," "veteran," "owned," and "Military" when discussing veterans. These terms emphasize the honor, service, and historical contributions of veterans, reflecting a focus on their roles in the military and their achievements. Overall, the results align with my expectations, showing that Democrats emphasize support and assistance for veterans, while Republicans highlight their service and contributions to the military.

Bonus Task

For Democrats

```
              term similarity rank
50          caused  0.3370485    1
49            says  0.3372709    2
48    humanitarian  0.3385240    3
47      associated  0.3395371    4
46          hiding  0.3404972    5
45      cigarettes  0.3410668    6
44    dramatically  0.3411431    7
43             cap  0.3413650    8
42          pushed  0.3414495    9
41         obvious  0.3419437   10
40      threatened  0.3444061   11
39             yet  0.3453570   12
38          rushed  0.3454467   13
37         morally  0.3458446   14
36          Abbott  0.3463349   15
35       Trumpcare  0.3466754   16
34         billion  0.3470794   17
33        decision  0.3477500   18
32        attempts  0.3488499   19
31         effects  0.3492332   20
30             Yet  0.3496993   21
29          passes  0.3497231   22
28          latest  0.3517745   23
27        behavior  0.3519297   24
26        possibly  0.3525255   25
25           GOP's  0.3532482   26
24            nail  0.3541245   27
23       proposing  0.3554061   28
22         version  0.3562774   29
```

21	repeal	0.3575486	30
20	direct	0.3583585	31
19	Administration's	0.3592943	32
18	voter	0.3598202	33
17	removing	0.3612635	34
16	Ban	0.3614123	35
15	kicking	0.3637095	36
14	skyrocket	0.3654003	37
13	kind	0.3673960	38
12	revealed	0.3702582	39
11	CBO	0.3702710	40
10	highly	0.3708080	41
9	Without	0.3755428	42
8	TrumpCare	0.3767717	43
7	withdraw	0.3777378	44
6	proposal	0.3803783	45
5	ram	0.3836270	46
4	raises	0.3836629	47
3	strip	0.4008645	48
2	attempt	0.4147846	49
1	result	0.4198848	50

For Republicans

	term	similarity	rank
50	terrorists	0.3159275	1
49	fires	0.3163150	2
48	spread	0.3167615	3
47	deductibles	0.3179511	4
46	type	0.3183182	5
45	car	0.3190796	6
44	engaged	0.3193706	7
43	commission	0.3193928	8
42	response	0.3201384	9
41	necessity	0.3202355	10
40	hurricanes	0.3202551	11
39	results	0.3227432	12
38	condition	0.3228828	13
37	Venezuela	0.3236159	14
36	deficit	0.3253799	15
35	Syrian	0.3255361	16
34	scrutiny	0.3264854	17

33	Iranian	0.3269852	18
32	pets	0.3270889	19
31	rise	0.3277351	20
30	government's	0.3291244	21
29	times	0.3291358	22
28	database	0.3292294	23
27	certain	0.3294412	24
26	threatens	0.3323452	25
25	overly	0.3332352	26
24	arrest	0.3332793	27
23	influence	0.3335205	28
22	settlements	0.3358009	29
21	privacy	0.3367529	30
20	crimes	0.3372485	31
19	sanctuary	0.3372520	32
18	warrant	0.3389339	33
17	requirement	0.3397620	34
16	wildfires	0.3403779	35
15	sexual	0.3415138	36
14	mandate	0.3420366	37
13	act	0.3441021	38
12	destroyed	0.3454546	39
11	Allowing	0.3462679	40
10	wildfire	0.3466651	41
9	ruling	0.3492462	42
8	probation	0.3511641	43
7	Instead	0.3522020	44
6	innocent	0.3543351	45
5	rules	0.3552207	46
4	accept	0.3567237	47
3	requiring	0.3586363	48
2	terrorist	0.3590129	49
1	powers	0.3639761	50

The results show some clear differences in the words that have the strongest negative associations for Democrats and Republicans. For Democrats, words like TrumpCare, repeal, replacement, deficit, and subsidies appear frequently. These terms reflect concerns about healthcare reforms and economic issues, such as Republican efforts to repeal the Affordable Care Act and worries about budget cuts and deficits.

For Republicans, words like sanctuary, terrorist, government's, and rules are more common. These words focus on issues like national security, government regulations, and law enforcement. Words like wildfires, violence, and victims suggest concern about natural disasters and personal safety.

When we look at the rankings of these words, we can see that both parties share some terms, such as deficit, but they likely use them differently. For Democrats, deficit may be linked to concerns about spending cuts, while Republicans may refer to it in the context of government overspending. Other shared words like cover, billion, and attempt appear for both parties, but the ranking of these words and their context differ.

Overall, the rankings reflect each party's priorities, with Democrats focusing more on healthcare and social policy, while Republicans are more concerned about security.

Bonus 7

I tried but my code did not work (though it is not obligatory)