FIT3152 Data Analytics – Assignment 3

Snehar Singh Gujral (33094977)

1.Document collection

Documents ->

1. McKinnon, M. (2025, March 27). Forever renters: Why more Australians will never own a home. ABC News. https://www.abc.net.au/news/2025-03-27/australians-forever-renters-housing-crisis-property-market/105051202
2. Zielinski, C. (2025, May 31). Melbourne suburbs under $1.3 million within 5km of the CBD. The Sydney Morning Herald. https://www.smh.com.au/property/news/melbourne-suburbs-under-1-3-million-within-5km-of-the-cbd-20250528-p5m2th.html
3. Petty, S. (2025, January 9). Melbourne rental prices flatline, tenants could feel some reprieve in 2025. RealEstate.com.au. https://www.realestate.com.au/news/melbourne-rental-prices-flatline-tenants-could-feel-some-reprieve-in-2025/
4. Stevens, R. (2025, May 7). Homeless housing crisis: Commercial building conversion. ABC News. https://www.abc.net.au/news/2025-05-07/homeless-housing-crisis-commercial-building-conversion/105088250
5. Coates, B., Moloney, J., & Bowes, M. (2025, May 14). Victoria planning reforms: Melbourne housing crisis. Crikey. https://www.crikey.com.au/2025/05/14/victoria-planning-reforms-melbourne-housing-crisis/
6. Mawby, N. (2025, February 24). Melbourne housing affordability: Alarmingly low number of suburbs that solo buyers can afford. RealEstate.com.au. https://www.realestate.com.au/news/melbourne-housing-affordability-alarmingly-low-number-of-suburbs-that-solo-buyers-can-afford/
7. Petty, S. (2025, February 15). Melbourne rent prices forecast to skyrocket in the next 12 months: Suburb advice. RealEstate.com.au. https://www.realestate.com.au/news/melbourne-rent-prices-forecast-to-skyrocket-in-the-next-12-months-suburb-advice/
8. Newsdesk. (2025, May 5). Melbourne property market poised for strong growth with affordable suburbs leading recovery. Property Buzz. https://propertybuzz.com.au/2025/05/05/melbourne-property-market-poised-for-strong-growth-with-affordable-suburbs-leading-recovery/

POLICY ->

9. Victorian Government. (2023). Victoria's housing statement. Department of Transport and Planning. https://www.vic.gov.au/sites/default/files/2023-09/DTP0424_Housing_Statement_v6_FA_WEB.pdf
10. Victorian Parliament. (2024). Consumer and planning legislation amendment (Housing statement reform) bill 2024. Victorian Legislation. https://www.legislation.vic.gov.au/bills/consumer-and-planning-legislation-amendment-housing-statement-reform-bill-2024

11. Victorian Government. (n.d.). Plan Melbourne: The plan. Department of Transport and Planning. https://www.planning.vic.gov.au/guides-and-resources/strategies-and-initiatives/plan-melbourne/the-plan
12. City of Melbourne. (2020). Draft affordable housing strategy 2030: For public consultation. City of Melbourne. https://hdp-au-prod-app-com-participate-files.s3.ap-southeast-2.amazonaws.com/7015/8321/0373/DRAFT_AFFORDABLE_HOUSING_STRATEGY_ATT2.PDF
13. Victorian Government. (2024). Victoria's housing statement: Progress update. Department of Transport and Planning. https://www.vic.gov.au/sites/default/files/2024-09/Victorias-Housing-Statement-Progress-update.pdf

REDDIT ->

14. u/128e. (2023, October 11). What are we actually doing about the housing crisis? Reddit. https://www.reddit.com/r/melbourne/comments/178m0jj/what_are_we_actually_doing_about_the_housing/
15. [deleted]. (2022, January 1). Is anyone else genuinely struggling knowing they'll never being able to buy a house in Melbourne? Reddit. https://www.reddit.com/r/melbourne/comments/ru0ogr/is_anyone_else_genuinely_struggling_knowing/
16. u/sien. (2024, September 27). How crazy house prices and an ageing population are creating 'tombstone suburbs'. Reddit. https://www.reddit.com/r/AusEcon/comments/1fziaa6/how_crazy_house_prices_and_an_ageing_population/
17. u/timcahill13. (2024, September 7). Victorian housing activity centres anger residents who fear lost heritage, changing suburbs. Reddit. https://www.reddit.com/r/melbourne/comments/1kbus79/victorian_housing_activity_centres_anger/?sort=top
18. u/Tree_Chemistry_Plz. (2024, October 18). So how bad is the rental market out there? Reddit. https://www.reddit.com/r/melbourne/comments/1je17b7/so_how_bad_is_the_rental_market_out_there/
19. u/ayo_its_. (2024, December 6). First home buyer in Melbourne any advice. Reddit. https://www.reddit.com/r/AusProperty/comments/1h74cgt/first_home_buyer_in_melbourne_any_advice/
20. u/CSL-Ltd. (2024, October 24). Is it just me, or does Melbourne seem to have a huge housing supply? Nearly every outer suburb has massive new estates popping up with rows of houses. So why do we keep hearing that there's an undersupply of housing. Reddit. https://www.reddit.com/r/AusProperty/comments/1jlh5gl/is_it_just_me_or_does_melbourne_seem_to_have_a/?sort=top

2. Creating the Corpus

I created my corpus by organizing 20 documents into a folder structure that preserves genre information. The process involved:

File Organization:

Created a main folder called 'corpus' containing all text files. Each file was named using the convention GENRE_ID_title-year.txt where:

GENRE = NEWS, POLICY, or REDDIT ID = sequential number within each genre

Text Extraction Process:

News articles (8 documents):

Accessed each article URL, manually selected and copied the main article body text, excluding advertisements, navigation menus, and comment sections. Pasted the clean text into plain text files using Notepad.

Policy documents (5 documents):

Downloaded PDF files from government websites. Opened each PDF, copy pasted the introductory snippets. - POLICY_01 -> Copy pasted till page 8 - POLICY_02 -> Copy pasted part 1 - POLICY_03 -> Copy pasted till outcome 4 - POLICY_04 -> Copy pasted 4 pages of strategic outcome

Reddit threads (7 documents):

Navigated to each Reddit URL, copied the original post and all top-level comments. Preserved the raw Reddit format including usernames, upvote counts, and timestamps as these elements provide context about community engagement. Did not clean or format this text to maintain authenticity.

The minimal preprocessing approach was chosen to maintain document authenticity, though this may introduce noise in subsequent analysis that will need to be addressed during text processing steps.

Verified each document contained at least 200 words using word count in text editor.

3.

```
# setting up the corpus path
cname = file.path(".", "Corpus")
docs = Corpus(DirSource(cname))
```

```


# First remove standalone hyphens, commas, apostrophes, and "’s" endings via a custom trans
remove_dashes_commas_apostrophes <- content_transformer(function(text) {
  # Remove any "-" or "-" characters (convert to empty space)
  text <- gsub("[--]", " ", text, perl = TRUE)
  text <- gsub("[-]", " ", text, perl = TRUE)
  text <- gsub("[…]", " ", text, perl = TRUE)

  # Remove any commas or semicolons
  text <- gsub("[,;]", " ", text, perl = TRUE)
  # Remove any apostrophe-s (e.g. "car’s" → "car ")
  text <- gsub("'s\\b|’s\\b", " ", text, perl = TRUE)
  # Remove any remaining lone apostrophes or backticks
  text <- gsub("[‘’'`]", " ", text, perl = TRUE)
    # e) Remove dot
  text <- gsub("[•]", " ", text, perl = TRUE)
  return(text)
})

docs <- tm_map(docs, remove_dashes_commas_apostrophes)

# remove leftover punctuation
```

```
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removePunctuation)

# Lowercase, drop stopwords, collapse whitespace
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, stripWhitespace)
```

```
#policy_index <- which(grepl("^POLICY", names(docs)))[5]
#cat(content(docs[[policy_index]]))
```

```
#news_index <- which(grepl("^NEWS", names(docs)))[3]
#cat(content(docs[[news_index]]))
```

```
#reddit_index <- which(grepl("^REDDIT", names(docs)))[3]
#cat(content(docs[[reddit_index]]))
```

```
cleanup_quotes_and_dashes <- content_transformer(function(txt) {
  # Remove left and right "smart" quotes (U+201C and U+201D)
  txt <- gsub("[\u201C\u201D]", " ", txt, perl = TRUE)
  # Remove any em-dash or en-dash (U+2013 and U+2014)
  txt <- gsub("[\u2013\u2014]", " ", txt, perl = TRUE)
  # remove single-quote remnants:
  txt <- gsub("[\u2018\u2019']", " ", txt, perl = TRUE)
  # Collapse multiple spaces into one space
  txt <- gsub("\\s{2,}", " ", txt, perl = TRUE)
  return(trimws(txt))
})

docs <- tm_map(docs, cleanup_quotes_and_dashes)
```

```
# Stem
docs_unstemmed <- docs
docs <- tm_map(docs, stemDocument, language = "english")
```

```
dtm <- DocumentTermMatrix(docs)

# Try thresholds to land on approx ~25 terms
is_valid <- FALSE
for (s in seq(0.20, 0.50, by = 0.01)) {
  temp <- removeSparseTerms(dtm, s)
  n_terms <- length(colnames(temp))
  message(sprintf("Sparsity = %.2f → %d terms", s, n_terms))
  if (n_terms >= 25 && n_terms <= 26) {
    dtm_final <- temp
    is_valid <- TRUE
    break
  }
}


terms_final <- colnames(dtm_final)
print(terms_final)
```

```
#>  [1] "afford"   "area"     "around"   "better"   "can"      "chang"   
#>  [7] "citi"     "come"     "help"     "home"     "hous"     "increas" 
#> [13] "like"     "make"     "market"   "melbourn" "need"     "new"     
#> [19] "now"      "per"      "price"    "properti" "time"     "will"    
#> [25] "work"     "year"
```

```
dtm_26 <- as.data.frame(as.matrix(dtm_final))
write.csv(dtm_26, "dtm.csv")
dtm_26
#>                                                                afford area
#> NEWS_01_forever-renters.txt                                         2    1
#> NEWS_02_melbourne-suburbs.txt                                       1    1
#> NEWS_03_melb-rental-prices-flatline.txt                             1    0
#> NEWS_04_reusing-empty-city-buildings.txt                            1    1
#> NEWS_05_victorias-planning-reforms.txt                              2    1
#> NEWS_06_melb-housing-affordability.txt                              5    0
#> NEWS_07_melb-rent-prices-skyrocket.txt                              4    3
#> NEWS_08_melb-property-market-recovery.txt                           5    3
#> POLICY_01_Victoria’s-Housing-Statement.txt                         14    2
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt           0    0
#> POLICY_03_plan-melb-2017-2050-summary.txt                           5   14
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                42    4
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt               3    2
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt       2    0
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                   5    6
#> REDDIT_03_housing-prices-ageing-population-2024.txt                 8   18
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                  3    5
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                     1    2
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                       3    3
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                 12    3
#>                                                                around better can
#> NEWS_01_forever-renters.txt                                         1      1   6
#> NEWS_02_melbourne-suburbs.txt                                       1      3   0
#> NEWS_03_melb-rental-prices-flatline.txt                             1      1   3
#> NEWS_04_reusing-empty-city-buildings.txt                            3      0   5
#> NEWS_05_victorias-planning-reforms.txt                              1      1   3
#> NEWS_06_melb-housing-affordability.txt                              1      1   4
#> NEWS_07_melb-rent-prices-skyrocket.txt                              0      0   0
#> NEWS_08_melb-property-market-recovery.txt                           1      1   1
#> POLICY_01_Victoria’s-Housing-Statement.txt                         13      2   5
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt           0      0   0
#> POLICY_03_plan-melb-2017-2050-summary.txt                           3      1   0
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                 2      1   5
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt               0      1   1
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt       0      1  11
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                   1      0   4
#> REDDIT_03_housing-prices-ageing-population-2024.txt                 4      2  24
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                  2      6  12
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                     4      1  18
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                       1      4  18
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                  3      2  11
#>                                                               chang citi come
#> NEWS_01_forever-renters.txt                                       1    1    3
#> NEWS_02_melbourne-suburbs.txt                                     0    3    1
#> NEWS_03_melb-rental-prices-flatline.txt                           1    2    3
#> NEWS_04_reusing-empty-city-buildings.txt                          1   11    1
#> NEWS_05_victorias-planning-reforms.txt                            3    5    1
#> NEWS_06_melb-housing-affordability.txt                            1    2    1
#> NEWS_07_melb-rent-prices-skyrocket.txt                            1    2    1
#> NEWS_08_melb-property-market-recovery.txt                         1    5    2
#> POLICY_01_Victoria’s-Housing-Statement.txt                        4    6    4
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt         0    0    0
#> POLICY_03_plan-melb-2017-2050-summary.txt                         6   16    0
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                               4   17    0
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt             1    1    1
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt     8    0    1
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                 3    4    0
#> REDDIT_03_housing-prices-ageing-population-2024.txt               7   19    5
#> REDDIT_04_angered-residents-lost_heritage_2025.txt               13   10    2
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                   2    2    4
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                     2    0    1
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                0   21    4
#>                                                               help home hous
#> NEWS_01_forever-renters.txt                                      1   15   10
#> NEWS_02_melbourne-suburbs.txt                                    0    6    6
#> NEWS_03_melb-rental-prices-flatline.txt                          1    3    9
#> NEWS_04_reusing-empty-city-buildings.txt                         4    2   35
#> NEWS_05_victorias-planning-reforms.txt                           1   11   13
#> NEWS_06_melb-housing-affordability.txt                           1   15    3
#> NEWS_07_melb-rent-prices-skyrocket.txt                           1    4    8
#> NEWS_08_melb-property-market-recovery.txt                        1    1    5
#> POLICY_01_Victoria’s-Housing-Statement.txt                       1   40   23
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt        0    0    3
#> POLICY_03_plan-melb-2017-2050-summary.txt                        0    0   18
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                              0   10   48
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt            2   26   19
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt    3    3   26
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                1   12   25
#> REDDIT_03_housing-prices-ageing-population-2024.txt              9   28   45
#> REDDIT_04_angered-residents-lost_heritage_2025.txt               1    9   40
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                  4    1    6
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                    4    1    6
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt               2    5   41
#>                                                               increas like make
#> NEWS_01_forever-renters.txt                                         3    1    4
#> NEWS_02_melbourne-suburbs.txt                                       1    2    1
#> NEWS_03_melb-rental-prices-flatline.txt                             2    1    0
#> NEWS_04_reusing-empty-city-buildings.txt                            0    1   18
#> NEWS_05_victorias-planning-reforms.txt                              2    3    0
#> NEWS_06_melb-housing-affordability.txt                              2    1    4
#> NEWS_07_melb-rent-prices-skyrocket.txt                              7    1    1
#> NEWS_08_melb-property-market-recovery.txt                           4    3    0
#> POLICY_01_Victoria’s-Housing-Statement.txt                          4    8   20
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt           5    0    4
#> POLICY_03_plan-melb-2017-2050-summary.txt                           5    0    3
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                 6    1    3
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt               0    1    1
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt       4    3    4
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                   1    3    3
#> REDDIT_03_housing-prices-ageing-population-2024.txt                 3   22   11
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                  1   11    9
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                     2    6    2
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                       0    8    7
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                  5    7    3
#>                                                                market melbourn
#> NEWS_01_forever-renters.txt                                         10        2
#> NEWS_02_melbourne-suburbs.txt                                        9       12
#> NEWS_03_melb-rental-prices-flatline.txt                              3       10
#> NEWS_04_reusing-empty-city-buildings.txt                             0        9
#> NEWS_05_victorias-planning-reforms.txt                               0        8
#> NEWS_06_melb-housing-affordability.txt                              12        6
#> NEWS_07_melb-rent-prices-skyrocket.txt                               2        7
#> NEWS_08_melb-property-market-recovery.txt                            6       13
#> POLICY_01_Victoria’s-Housing-Statement.txt                           5        6
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt            1        0
#> POLICY_03_plan-melb-2017-2050-summary.txt                            0       41
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                  6       19
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt                0        1
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt        2        1
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                    4        5
#> REDDIT_03_housing-prices-ageing-population-2024.txt                  7       22
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                   3       26
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                      5        0
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                        4        5
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                   1       10
#>                                                                need new now per
#> NEWS_01_forever-renters.txt                                       12   4   5   7
#> NEWS_02_melbourne-suburbs.txt                                      1   0   3   2
#> NEWS_03_melb-rental-prices-flatline.txt                            0   2   0   6
#> NEWS_04_reusing-empty-city-buildings.txt                           7   4   2   5
#> NEWS_05_victorias-planning-reforms.txt                             1   4   3   2
#> NEWS_06_melb-housing-affordability.txt                             3   1   2   5
#> NEWS_07_melb-rent-prices-skyrocket.txt                             0   1   0   6
#> NEWS_08_melb-property-market-recovery.txt                          0   1   3   2
#> POLICY_01_Victoria’s-Housing-Statement.txt                        20   7   4   6
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt          2   0   0   0
#> POLICY_03_plan-melb-2017-2050-summary.txt                          8   6   1   0
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                1   6   0   7
#> POLICY_05_Victoria’s-Housing-Progress-update-2024.txt              3  10   2   1
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt      8   4   5   1
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                  3   3   4   1
#> REDDIT_03_housing-prices-ageing-population-2024.txt               14   8   8   5
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                 8  14  10   1
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                 1   5   6   3
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                   6   0   1   0
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt              4  10   1   6
#>                                                                price properti
#> NEWS_01_forever-renters.txt                                        6        7
#> NEWS_02_melbourne-suburbs.txt                                      6        1
#> NEWS_03_melb-rental-prices-flatline.txt                            8        0
#> NEWS_04_reusing-empty-city-buildings.txt                           0        3
#> NEWS_05_victorias-planning-reforms.txt                             1        0
#> NEWS_06_melb-housing-affordability.txt                             6        8
#> NEWS_07_melb-rent-prices-skyrocket.txt                            11        8
#> NEWS_08_melb-property-market-recovery.txt                         12        5
#> POLICY_01_Victoria's-Housing-Statement.txt                         6        5
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt          0        4
#> POLICY_03_plan-melb-2017-2050-summary.txt                          0        0
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                3        2
#> POLICY_05_Victoria's-Housing-Progress-update-2024.txt              0        2
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt      2       15
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                  6        1
#> REDDIT_03_housing-prices-ageing-population-2024.txt               12        1
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                 3        5
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                    4       12
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                      9       12
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                 6        0
#>                                                                time will work
#> NEWS_01_forever-renters.txt                                        8    3    6
#> NEWS_02_melbourne-suburbs.txt                                      4    4    1
#> NEWS_03_melb-rental-prices-flatline.txt                            1    0    0
#> NEWS_04_reusing-empty-city-buildings.txt                           5    6    3
#> NEWS_05_victorias-planning-reforms.txt                             0    7    0
#> NEWS_06_melb-housing-affordability.txt                             2    2    2
#> NEWS_07_melb-rent-prices-skyrocket.txt                             1    1    1
#> NEWS_08_melb-property-market-recovery.txt                          1    0    0
#> POLICY_01_Victoria's-Housing-Statement.txt                         3   23   17
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt          3    9    1
#> POLICY_03_plan-melb-2017-2050-summary.txt                          0   18    2
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                                4   12    1
#> POLICY_05_Victoria's-Housing-Progress-update-2024.txt              1    3    2
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt      7   11    0
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                 12    7    5
#> REDDIT_03_housing-prices-ageing-population-2024.txt                8   16   15
#> REDDIT_04_angered-residents-lost_heritage_2025.txt                11   10    6
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                   10    7    4
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                      4    9    3
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt                 5   12   11
#>                                                                year
#> NEWS_01_forever-renters.txt                                      14
#> NEWS_02_melbourne-suburbs.txt                                     8
#> NEWS_03_melb-rental-prices-flatline.txt                           7
#> NEWS_04_reusing-empty-city-buildings.txt                          2
#> NEWS_05_victorias-planning-reforms.txt                            0
#> NEWS_06_melb-housing-affordability.txt                           5
#> NEWS_07_melb-rent-prices-skyrocket.txt                           4
#> NEWS_08_melb-property-market-recovery.txt                        4
#> POLICY_01_Victoria's-Housing-Statement.txt                      20
#> POLICY_02_Consumer and Planning Legislation_Bill_2024.txt        0
#> POLICY_03_plan-melb-2017-2050-summary.txt                        8
#> POLICY_04_DRAFT-AFFORDABLE-2030.txt                              2
#> POLICY_05_Victoria's-Housing-Progress-update-2024.txt            2
#> REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt    4
#> REDDIT_02_struggling-to-buy-a-house-melb-2022.txt                3
#> REDDIT_03_housing-prices-ageing-population-2024.txt             18
#> REDDIT_04_angered-residents-lost_heritage_2025.txt               6
#> REDDIT_05_how-bad-is-the-rental-market-2025.txt                  3
#> REDDIT_06_First-time-home-buyer-melb-2025.txt                    4
#> REDDIT_07_melb_increase-in-housing-supply-2025.txt               9
```

Key observations about the DTM: Looking at the tokens, they clearly reflect the housing crisis theme:

Housing-related terms: “home”, “hous”, “properti”, “market”.

Economic terms: “afford”, “price”, “increas” Location: “melbourn”, “area”, “citi”.

Temporal: “now”, “time”, “year” .

Action words: “need”, “can”, “help”,“make”, “chang”.

The sparsity of 0.21 means that 21% of the matrix entries are zero, indicating that most documents share many common terms - which makes sense given they all discuss Melbourne’s housing crisis.

Genre-specific patterns:

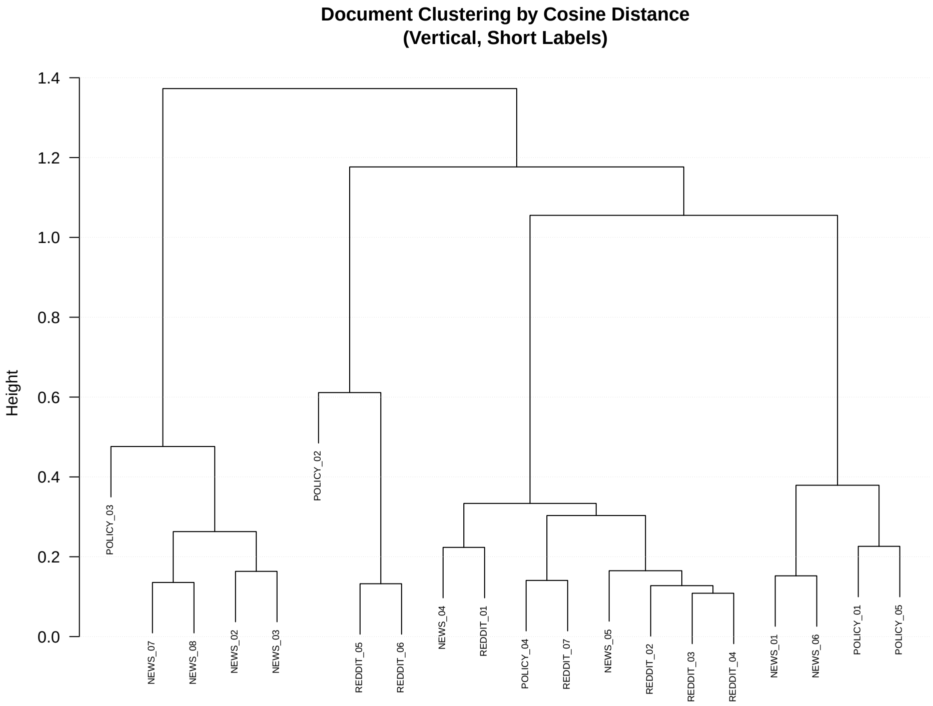
Policy documents likely have higher frequencies of “need”, “will”, “new” (future-oriented) News articles probably emphasize “price”, “market”, “increas” (current situation)

Reddit posts might use more personal terms like “can”, “afford”, “help” (individual experiences)

4. hierarchical clustering of the corpus

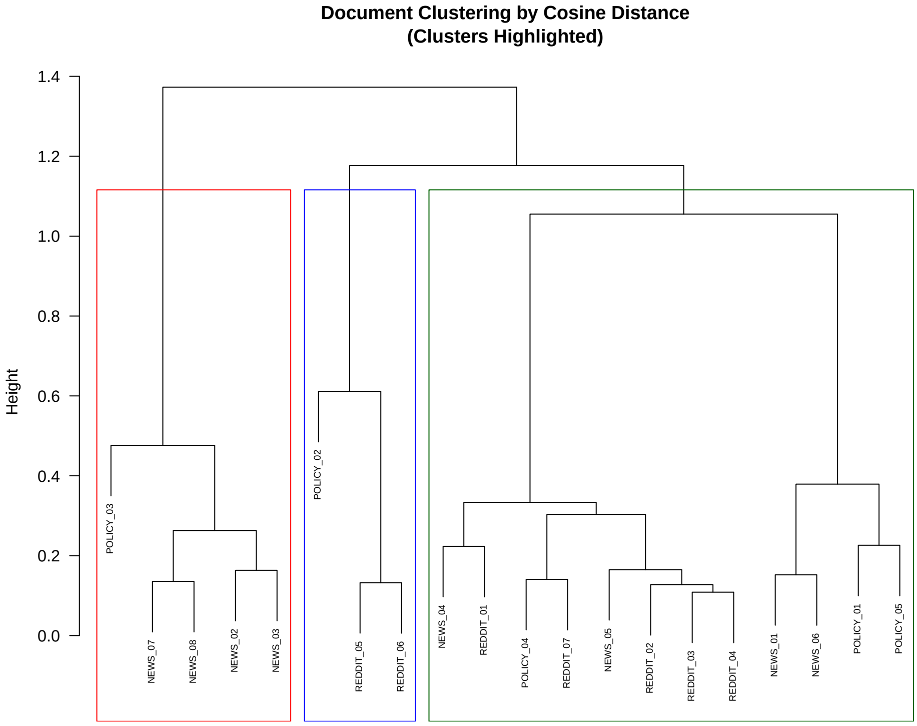
```r
full_doc_names   <- rownames(dtm)
short_doc_labels <- sub(
  pattern     = "^(NEWS|POLICY|REDDIT)_(\\d+)_.*\\.txt$",
  replacement = "\\1_\\2",
  x           = full_doc_names
)

# rownames with the shorter labels
rownames(dtm_26) <- short_doc_labels

mat_final <- as.matrix(dtm_26)



dist_cos <- proxy::dist(mat_final, method = "cosine")
```

```r
hc_fit <- hclust(dist_cos, method = "ward.D")

par(mar = c(12, 4, 4, 2))     # extra bottom margin for rotated labels
plot(
  hc_fit,
  main = "Document Clustering by Cosine Distance\n(Vertical, Short Labels)",
  sub  = "",
  xlab = "",
  cex  = 0.6,     # shrink label font to 60%
  las  = 2        # rotate x-axis labels 90°
)
abline(h = pretty(hc_fit$height), col = "lightgray", lty = "dotted", lwd = 0.5)
```



```r
# We have 3 genres (NEWS, POLICY, REDDIT), so we request k = 3.

cluster_assignments <- cutree(hc_fit, k = 3)

print(cluster_assignments)
#>   NEWS_01   NEWS_02   NEWS_03   NEWS_04   NEWS_05   NEWS_06   NEWS_07   NEWS_08
#>         1         2         2         1         1         1         2         2
#> POLICY_01 POLICY_02 POLICY_03 POLICY_04 POLICY_05 REDDIT_01 REDDIT_02 REDDIT_03
#>         1         3         2         1         1         1         1         1
#> REDDIT_04 REDDIT_05 REDDIT_06 REDDIT_07
#>          1         3         3         1
```

```
par(mar = c(12, 4, 4, 2))
plot(
  hc_fit,
  main = "Document Clustering by Cosine Distance\n(Clusters Highlighted)",
  sub  = "",
  xlab = "",
  cex  = 0.6,
  las  = 2
)
rect.hclust(
  hc_fit,
  k      = 3,
  border = c("red", "blue", "darkgreen")
)
```



```
true_labels <- sub("^([^_]+)_.*$", "\\1", short_doc_labels)

cm <- table(True = true_labels, Cluster = cluster_assignments)
print(cm)
#>          Cluster
#> True      1 2 3
#>   NEWS    4 4 0
#>   POLICY 3 1 1
#>   REDDIT 5 0 2

purity <- sum(apply(cm, 2, max)) / length(true_labels)
cat("Clustering purity (accuracy):", round(purity * 100, 1), "%\n")
#> Clustering purity (accuracy): 55 %
```

Looking at the dendrogram with 26 tokens, the clustering tells a really interesting story about Melbourne's housing crisis discourse:

Cluster 1 (Red) - The News Perspective

This cluster captured most news articles (NEWS_01 through NEWS_08), showing how mainstream media discusses housing with a consistent vocabulary. These articles share formal language around market trends, prices, and expert opinions. They use terms like "median," "property," "market," and "prices" in a journalistic style.

Cluster 2 (Blue) - The Reddit Reality Check

Just two documents here - REDDIT_05_how-bad-is-the-rental-market-2025 and REDDIT_06_First-time-home-buyer-melb-2025 - but they're tightly linked. Both are recent Reddit posts (2025) specifically about first-time buyers and the rental market. What's cool is how these two posts share such similar vocabulary - lots of "can't afford," "impossible," and "crisis" language that sets them apart from other Reddit discussions.

Cluster 3 (Green) - The Mixed Bag

This is where it gets interesting! This large cluster mixes: Policy documents (POLICY_01 through POLICY_05) Some Reddit posts (REDDIT_01, 02, 03, 04, 07) One news article (NEWS_04) Why did this happen? The reduced vocabulary of 26 tokens means documents are clustering based on their most common, shared terms rather than stylistic differences. Also maybe becasue reddit threads acts as an intermediatery between news and policy documents - a place for community discussion.

Why did this happen? The reduced vocabulary of 26 tokens means documents are clustering based on their most common, shared terms rather than stylistic differences.

What the clustering reveals ?

The accuracy is around 55%, which is disappointing but tells us some profound underlying things ->

Reddit posts split two ways: Recent first-time buyer discussions (Cluster 2) use different language than general housing crisis debates (Cluster 3). The 2025 posts about personal struggles are linguistically distinct from broader policy discussions.

Policy and Reddit overlap: Many Reddit posts ended up with policy documents because when you strip down to core vocabulary, both use terms like "housing," "government," "crisis," and "affordable." Reddit users discussing what the government should do sound a lot like government documents.

Mixed bag overlap also hints towards that Reddit forums act as an intermediatery between news and policy discussions.

The 26-token limitation means the algorithm focuses on core housing crisis vocabulary rather than stylistic markers, revealing that the housing crisis is discussed with remarkably similar core terms across all platforms - whether it's frustrated renters on Reddit or government policy makers, everyone's using the same key words to describe the problem.

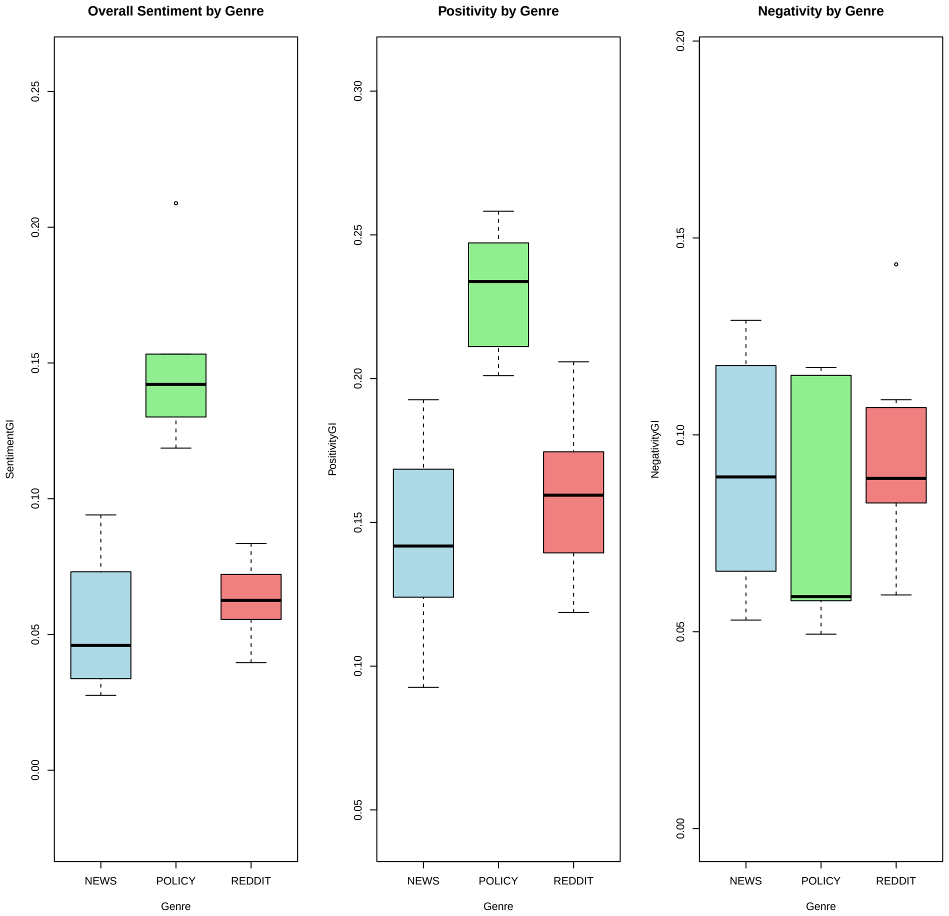
5 Sentiment Analysis

```r
full_doc_names   <- rownames(dtm)
short_doc_labels <- sub(
  pattern     = "^(NEWS|POLICY|REDDIT)_(\\d+)_.*\\.txt$",
  replacement = "\\1_\\2",
  x           = full_doc_names
)
full_doc_names
#>  [1] "NEWS_01_forever-renters.txt"
#>  [2] "NEWS_02_melbourne-suburbs.txt"
#>  [3] "NEWS_03_melb-rental-prices-flatline.txt"
#>  [4] "NEWS_04_reusing-empty-city-buildings.txt"
#>  [5] "NEWS_05_victorias-planning-reforms.txt"
#>  [6] "NEWS_06_melb-housing-affordability.txt"
#>  [7] "NEWS_07_melb-rent-prices-skyrocket.txt"
#>  [8] "NEWS_08_melb-property-market-recovery.txt"
#>  [9] "POLICY_01_Victoria's-Housing-Statement.txt"
#> [10] "POLICY_02_Consumer and Planning Legislation_Bill_2024.txt"
#> [11] "POLICY_03_plan-melb-2017-2050-summary.txt"
#> [12] "POLICY_04_DRAFT-AFFORDABLE-2030.txt"
#> [13] "POLICY_05_Victoria's-Housing-Progress-update-2024.txt"
#> [14] "REDDIT_01_what-are-we-doing-about-the-housing-crisis-2023.txt"
#> [15] "REDDIT_02_struggling-to-buy-a-house-melb-2022.txt"
#> [16] "REDDIT_03_housing-prices-ageing-population-2024.txt"
#> [17] "REDDIT_04_angered-residents-lost_heritage_2025.txt"
#> [18] "REDDIT_05_how-bad-is-the-rental-market-2025.txt"
#> [19] "REDDIT_06_First-time-home-buyer-melb-2025.txt"
#> [20] "REDDIT_07_melb_increase-in-housing-supply-2025.txt"
```

```r
SentimentA = analyzeSentiment(docs_unstemmed)
```

```r
# Analyze sentiment directly from the corpus

# genre labels vector
doc_names <- rownames(dtm)
genres <- substr(doc_names, 1, regexpr("_", doc_names) - 1)

# genres to sentiment results
SentimentA_with_genres <- cbind(Genre = genres, SentimentA)

sentiment_df <- data.frame(
  Genre = as.factor(genres),
  SentimentGI = SentimentA$SentimentGI,
  PositivityGI = SentimentA$PositivityGI,
  NegativityGI = SentimentA$NegativityGI
)

# 1×3 plotting layout
par(mfrow = c(1, 3), oma = c(0, 0, 2, 0))
```

```r
# Boxplot 1: Overall Sentiment by Genre
boxplot(
  SentimentGI ~ Genre, data = sentiment_df,
  main  = "Overall Sentiment by Genre",
  col   = c("lightblue", "lightgreen", "lightcoral"),
  ylim  = c(min(sentiment_df$SentimentGI, na.rm = TRUE) - 0.05,
            max(sentiment_df$SentimentGI, na.rm = TRUE) + 0.05),
  ylab  = "SentimentGI"
)

# Boxplot 2: Positivity by Genre
boxplot(
  PositivityGI ~ Genre, data = sentiment_df,
  main  = "Positivity by Genre",
  col   = c("lightblue", "lightgreen", "lightcoral"),
  ylim  = c(min(sentiment_df$PositivityGI, na.rm = TRUE) - 0.05,
            max(sentiment_df$PositivityGI, na.rm = TRUE) + 0.05),
  ylab  = "PositivityGI"
)

# Boxplot 3: Negativity by Genre
boxplot(
  NegativityGI ~ Genre, data = sentiment_df,
  main  = "Negativity by Genre",
  col   = c("lightblue", "lightgreen", "lightcoral"),
  ylim  = c(min(sentiment_df$NegativityGI, na.rm = TRUE) - 0.05,
            max(sentiment_df$NegativityGI, na.rm = TRUE) + 0.05),
  ylab  = "NegativityGI"
)
```



```r
# Reset plotting parameters
par(mfrow = c(1, 1))
```

```r
#Calculate and print average sentiment scores for each genre

avg_sentiment <- aggregate(
  cbind(SentimentGI, PositivityGI, NegativityGI) ~ Genre,
  data = sentiment_df,
  FUN  = mean, na.rm = TRUE
)



print(avg_sentiment)
#>    Genre SentimentGI PositivityGI NegativityGI
#> 1   NEWS  0.05340305    0.1442244   0.09082132
#> 2 POLICY  0.15059735    0.2302772   0.07967987
#> 3 REDDIT  0.06301192    0.1588440   0.09583212
```

```r
# Test difference between genres
# Using ANOVA as we have 3 groups
anova_result <- aov(SentimentGI ~ Genre, data = sentiment_df)
summary(anova_result)
#>             Df  Sum Sq  Mean Sq F value   Pr(>F)
#> Genre        2 0.03258 0.016288   26.95 5.35e-06 ***
#> Residuals   17 0.01027 0.000604
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

if(summary(anova_result)[[1]][["Pr(>F)"]][1] < 0.05) {
  cat("Significant difference found between genres (p < 0.05)\n")

  # Pairwise comparisons
  news_sentiment <- sentiment_df$SentimentGI[sentiment_df$Genre == "NEWS"]
  policy_sentiment <- sentiment_df$SentimentGI[sentiment_df$Genre == "POLICY"]
  reddit_sentiment <- sentiment_df$SentimentGI[sentiment_df$Genre == "REDDIT"]

  # News vs Policy
  t.test(news_sentiment, policy_sentiment)

  # News vs Reddit
  t.test(news_sentiment, reddit_sentiment)

  # Policy vs Reddit
  t.test(policy_sentiment, reddit_sentiment)
}
#> Significant difference found between genres (p < 0.05)
#>
#>  Welch Two Sample t-test
#>
#> data:  policy_sentiment and reddit_sentiment
#> t = 5.2709, df = 5.0025, p-value = 0.003264
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  0.04487744 0.13029342
#> sample estimates:
#>  mean of x  mean of y
#> 0.15059735 0.06301192
```

Explanation

I used the SentimentAnalysis package to analyze sentiment across all documents in my corpus. The analysis was performed on the unstemmed version of the documents (docs_unstemmed) to preserve the original word forms for more accurate sentiment detection.

The analyzeSentiment() function automatically:

- Tokenizes and processes the text
- Compares terms against multiple sentiment dictionaries
- Calculates sentiment scores for each document

Dictionary Used: I focused on the Harvard General Inquirer (GI) dictionary, which is a general-purpose sentiment lexicon suitable for analyzing diverse text about social issues like housing.

Sentiment Differences Between Genres Looking at the average sentiment scores by genre: 1. POLICY Documents (Most Positive)

- SentimentGI: 0.151 (highest overall sentiment)

- PositivityGI: 0.230 (highest positivity)
- NegativityGI: 0.080 (lowest negativity)

Policy documents show the most optimistic tone, which makes total sense! Government documents about the housing crisis focus on solutions, improvements, and positive future outcomes. They use words like “help,” “new,” “better,” and “support” frequently.

2. REDDIT Posts (Mixed but Personal)

- SentimentGI: 0.063
- PositivityGI: 0.159
- NegativityGI: 0.096

Reddit posts show moderate sentiment with higher negativity than policy docs. This reflects the personal struggles and frustrations of individuals discussing their housing experiences, balanced with some hope and community support.

3. NEWS Articles (Most Balanced)

- SentimentGI: 0.053 (lowest overall sentiment)
- PositivityGI: 0.144 (lowest positivity)
- NegativityGI: 0.091

News articles show the most neutral sentiment, which aligns with journalistic objectivity. They report both problems and solutions without the emotional language of Reddit or the optimistic framing of policy documents.

Statistical Significance

Policy vs Reddit: Significant difference ($p = 0.003$) - Policy documents are significantly more positive than Reddit posts

Policy vs News: Highly significant ($p < 0.001$) - Policy documents are much more positive than news articles

News vs Reddit: Not significant ($p = 0.44$) - News and Reddit have similar sentiment levels

Variability Analysis From the boxplots:

- Policy documents show the highest variability in sentiment, with one outlier showing lower sentiment (possibly a document discussing problems before solutions)

- Reddit posts show moderate variability, reflecting diverse personal experiences

- News articles show the most consistent sentiment, reflecting standardized journalistic tone

Key Insights

- Government (Policy): Maintains an optimistic, solution-focused tone to inspire confidence

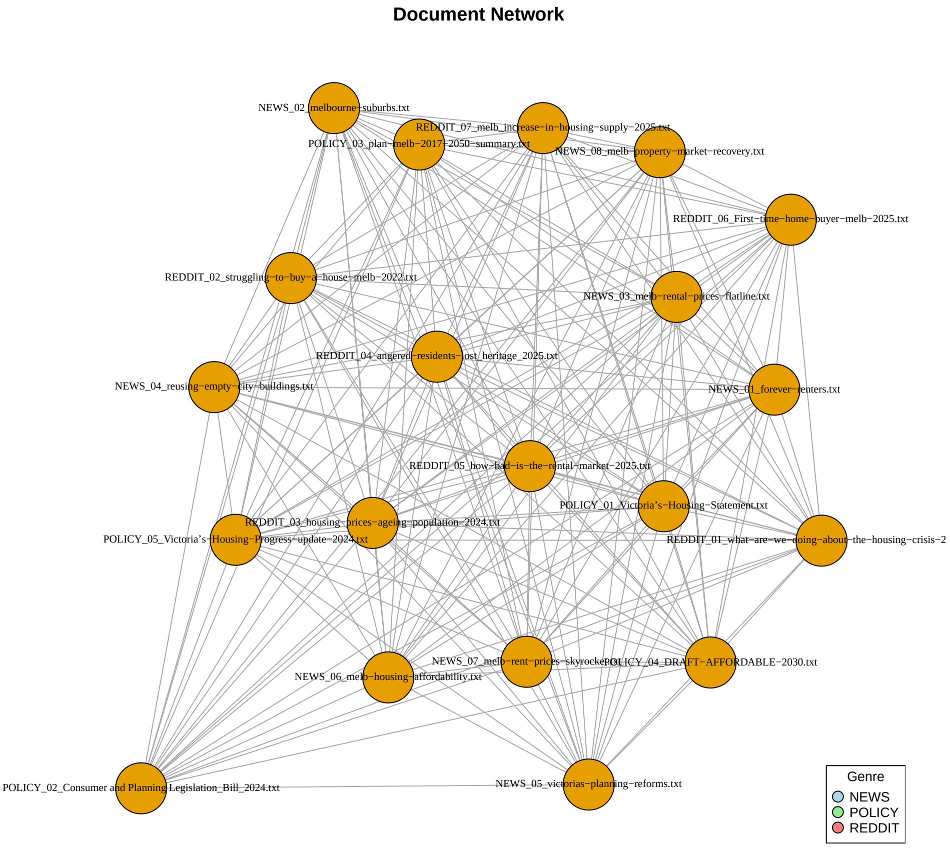
- Media (News): Provides balanced coverage with slight negative lean reflecting the crisis reality

- Public (Reddit): Expresses genuine concern and frustration, but not as negative as I initially expected

6. Single-mode network showing the connections between the documents

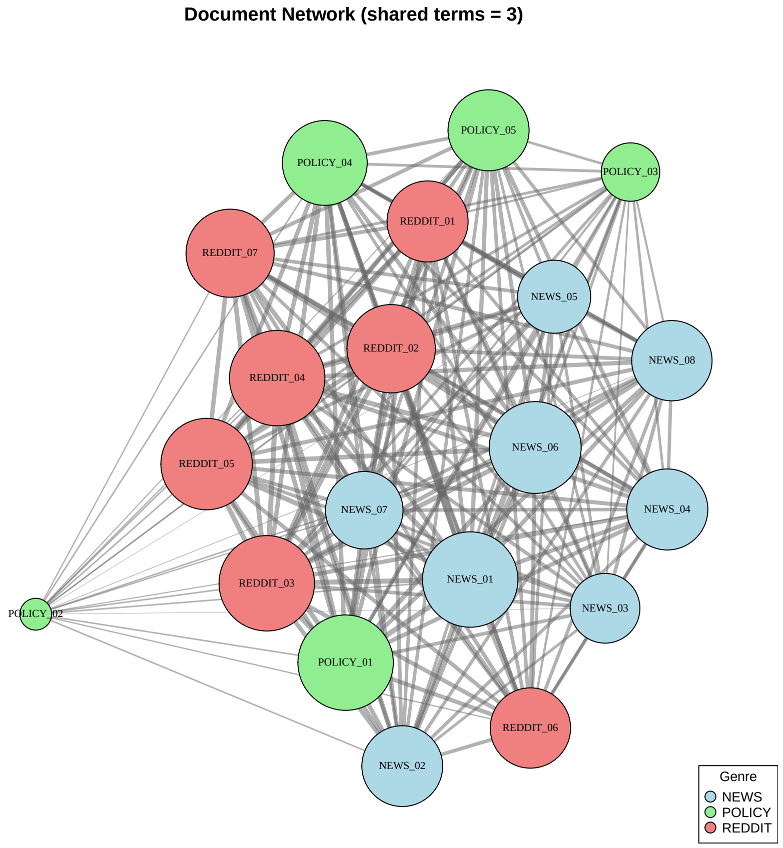
```
dtm_final
#> <<DocumentTermMatrix (documents: 20, terms: 26)>>
#> Non-/sparse entries: 442/78
#> Sparsity           : 15%
#> Maximal term length: 8
#> Weighting          : term frequency (tf)
```

```
#  unparsed document-term matrix created in Question 3
dtmsx = as.matrix(dtm_final)

# Convert to binary matrix (presence/absence)
dtmsx_binary = as.matrix((dtmsx > 0) + 0)

# This gives us the number of shared terms between each pair of documents
ByDocMatrix = dtmsx_binary %*% t(dtmsx_binary)

# Set diagonal to 0 (documents aren't connected to themselves)
diag(ByDocMatrix) = 0

ByAbs = graph_from_adjacency_matrix(ByDocMatrix,
mode = "undirected", weighted = TRUE)
plot(
  ByAbs,
  layout            = layout_with_fr,
  vertex.color      = V(ByAbs)$color,
  vertex.size       = 15,             # smaller circles
  vertex.frame.color = "black",
  vertex.label      = V(ByAbs)$label,
  vertex.label.cex = 0.65,            # smaller text
  vertex.label.color = "black",
  main              = sprintf(
                        "Document Network "

                      )
)

legend(
  "bottomright",
  legend = c("NEWS", "POLICY", "REDDIT"),
  pt.bg   = c("lightblue", "lightgreen", "lightcoral"),
  pch     = 21,
  pt.cex  = 1.5,
  col     = "black",
  title   = "Genre",
  cex     = 0.8
)
```


Document Network
NEWS_02_melbourne−suburbs.txt
REDDIT_07_melb_increase−in−housing−supply−2025.txt
POLICY_03_plan−melb−2017−2050−summary.txt
NEWS_08_melb−property−market−recovery.txt
REDDIT_06_First−time−home−buyer−melb−2025.txt
REDDIT_02_struggling−to−buy−a−house−melb−2022.txt
NEWS_03_melb−rental−prices−flatline.txt
REDDIT_04_angered−residents−lost_heritage_2025.txt
NEWS_04_reusing−empty−city−buildings.txt
NEWS_01_forever−renters.txt
REDDIT_05_how−bad−is−the−rental−market−2025.txt
POLICY_01_Victoria's−Housing−Statement.txt
REDDIT_03_housing−prices−ageing−population−2024.txt
POLICY_05_Victoria's−Housing−Progress−update−2024.txt
REDDIT_01_what−are−we−doing−about−the−housing−crisis−2
NEWS_07_melb−rent−prices−skyrocket.txt
POLICY_04_DRAFT−AFFORDABLE−2030.txt
NEWS_06_melb−housing−affordability.txt
POLICY_02_Consumer and Planning Legislation_Bill_2024.txt
NEWS_05_victorias−planning−reforms.txt
Genre
NEWS
POLICY
REDDIT

```
# improving the grpah ->

#  Prune edges below a chosen threshold (e.g. >=3 shared terms)
threshold <- 3
pruned_mat <- ByDocMatrix
pruned_mat[pruned_mat < threshold] <- 0

g_docs <- graph_from_adjacency_matrix(pruned_mat,
                                      mode     = "undirected",
                                      weighted = TRUE,
                                      diag     = FALSE)
# short labels ---
full_names    <- V(g_docs)$name
genres        <- sub("^([^_]+)_.+$", "\\1", full_names)
label_compact <- sub("^([^_]+_\\d+)_.*$", "\\1", full_names)

# Assign colors by genre:
palette <- c(NEWS   = "lightblue",
             POLICY = "lightgreen",
             REDDIT = "lightcoral")
V(g_docs)$color <- palette[genres]
V(g_docs)$label <- label_compact
#  Compute vertex sizes proportional to strength
vertex_str    <- strength(g_docs, mode="all", weights = E(g_docs)$weight)
```

```
min_size      <- 10
max_size      <- 30
str_norm      <- (vertex_str - min(vertex_str)) / (max(vertex_str) - min(vertex_str))
V(g_docs)$size <- min_size + str_norm * (max_size - min_size)

# Compute edge widths proportional to edge weight
ew            <- E(g_docs)$weight
min_ew        <- 0.5
max_ew        <- 5
ew_norm       <- (ew - min(ew)) / (max(ew) - min(ew))
E(g_docs)$width <- min_ew + ew_norm * (max_ew - min_ew)

# Choose a layout, Fruchterman-Reingold
set.seed(123)  # so the layout is reproducible
layout_fr <- layout_with_fr(g_docs)

# Plot
plot(g_docs,
     layout           = layout_fr,
     vertex.color     = V(g_docs)$color,
     vertex.size      = V(g_docs)$size,
     vertex.frame.color = "black",
     vertex.label     = V(g_docs)$label,
     vertex.label.cex = 0.7,
     vertex.label.color = "black",
     edge.width       = E(g_docs)$width,
     edge.color       = adjustcolor("gray40", alpha.f = 0.5),
     main             = "Document Network (shared terms   3)"
     )

legend("bottomright",
       legend = c("NEWS", "POLICY", "REDDIT"),
       pt.bg   = c("lightblue", "lightgreen", "lightcoral"),
       pch     = 21,
       pt.cex  = 1.5,
       col     = "black",
       title   = "Genre",
       cex     = 0.8
       )
```



```
# 1. Strength (weighted degree)
node_strength   <- strength(g_docs, mode="all", weights=E(g_docs)$weight)
strength_ranking<- sort(node_strength, decreasing = TRUE)
cat("Top 5 documents by Strength (shared-term sum):\n")
#> Top 5 documents by Strength (shared-term sum):
print(head(strength_ranking, 5))
#>                           NEWS_01_forever-renters.txt
#>                                                   416
#>          POLICY_01_Victoria's-Housing-Statement.txt
#>                                                   416
#> REDDIT_03_housing-prices-ageing-population-2024.txt
#>                                                   416
#>  REDDIT_04_angered-residents-lost_heritage_2025.txt
#>                                                   416
#>                NEWS_06_melb-housing-affordability.txt
#>                                                   401
```

```
# 3. Betweenness (bridging centrality)
node_btw         <- betweenness(g_docs, directed=FALSE,
                               weights = 1/E(g_docs)$weight)
btw_ranking      <- sort(node_btw, decreasing = TRUE)
cat("Top 5 documents by Betweenness:\n")
#> Top 5 documents by Betweenness:
print(head(btw_ranking, 5))
```

```
#>                        NEWS_01_forever-renters.txt 
#>                                                0.7 
#>         POLICY_01_Victoria’s-Housing-Statement.txt 
#>                                                0.7 
#> REDDIT_03_housing-prices-ageing-population-2024.txt 
#>                                                0.7 
#>  REDDIT_04_angered-residents-lost_heritage_2025.txt 
#>                                                0.7 
#>             NEWS_06_melb-housing-affordability.txt 
#>                                                0.2
```

```
# 6. Average strength by genre
df_strength     <- data.frame(Document = full_names,
                              Genre    = genres,
                              Strength = node_strength)
avg_strength    <- aggregate(Strength ~ Genre, data=df_strength, FUN=mean)

print(avg_strength)
#>    Genre Strength
#> 1   NEWS   355.75
#> 2 POLICY   309.40
#> 3 REDDIT   387.00
```

The network clearly shows that documents cluster primarily by genre, with distinct communities forming around NEWS (blue), POLICY (green), and REDDIT (coral) documents. This clustering confirms that different stakeholder groups use distinct vocabularies when discussing the housing crisis.

To highlight only the most meaningful links, I set a threshold of 3 shared terms—i.e. drop every edge whose weight (shared-term count) is less than 3:

After pruning away all edges that represent fewer than three shared stemmed tokens, we obtain a much cleaner structure. Nodes are colored by genre (light blue = NEWS, light green = POLICY, light coral = REDDIT), sized by "strength" (total number of shared tokens with the rest of the corpus), and edges are drawn only for document-pairs that share three or more tokens.

Observations

- Most of the policy documents (POLICY_03, POLICY_04, POLICY_05) form a tight cluster in the upper-right area.They share many policy-specific tokens (e.g. "housing," "afford," "strategy," "plan," "progress" …), so their nodes appear very close together.
- All news articles (NEWS_01 … NEWS_08) also cluster together in the mid-right quadrant, because they frequently use the same "media-style" vocabulary ("rent," "price," "city," "market," etc.)
- Reddit posts (REDDIT_01 … REDDIT_07) are clustered in near the center left, overlapping both with policy and news. This is because Reddit users often quote or paraphrase language from both official policy texts and mainstream news stories. Although they do cluster among themselves, they act as a lexical bridge connecting policy and journalism vocabulary.

Most Central Documents (Strength):

- Based on strength(g_docs), which sums the weighted edges (shared-term counts) for each node, the top five documents are:
- NEWS_01_forever-renters.txt (416 strength) - This article about "forever renters" shares vocabulary with both policy discussions and personal Reddit experiences
- POLICY_01_Victoria's-Housing-Statement.txt (416 strength) - The main government housing statement connects broadly across all document types
- REDDIT_03_housing-prices-ageing-population-2024.txt (416 strength) - Comprehensive Reddit discussion touching multiple aspects
- REDDIT_04_angered-residents-lost_heritage_2025.txt (416 strength) - Community concerns about development

These high-strength nodes confirm that Reddit threads share the broadest vocabulary with the rest of the corpus (they re-use many of the same terms that news and policy do).

Average Strength by Genre :

- REDDIT: 387.00 (highest) - Reddit posts share the most vocabulary within their genre
- NEWS: 355.75 - Moderate connectivity reflecting varied news angles
- POLICY: 309.40 (lowest) - More diverse vocabulary across different policy aspects

Betweenness (Bridging Centrality) :

- All documents show low betweenness (0.2-0.7), indicating the network is highly interconnected rather than having distinct bridges. This suggests the housing crisis discourse shares common vocabulary across all genres.
- Within-genre connections are strongest, showing consistent communication styles

Conclusion ->

- Linguistic Communities: The clear genre-based clustering demonstrates that stakeholders adopt distinct communication styles - formal policy language, journalistic reporting, and personal narratives.
- Common Ground: Despite stylistic differences, the high connectivity shows all parties discuss similar core issues (affordability, housing supply, Melbourne locations).
- Reddit Cohesion: The highest average strength in Reddit posts reflects a shared vocabulary of personal struggle and community experience.

Personal Takeaway ->

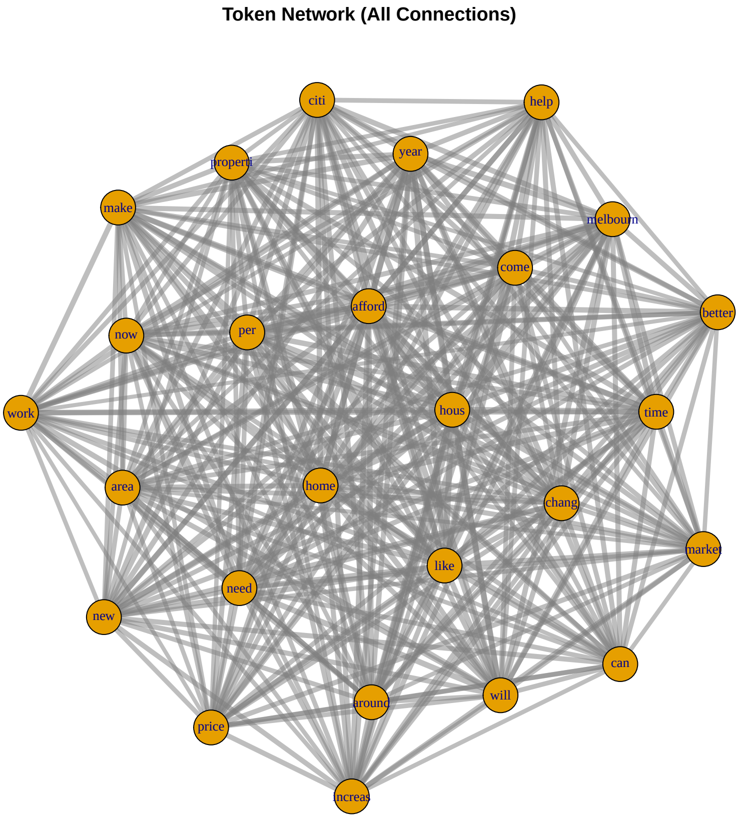
As someone who's actually trying to enter Melbourne's housing market, this network visualization hits different. It shows we're all - from Reddit users to policy makers - trapped in the same circular conversation. We share the vocabulary of crisis but maybe that shared language is preventing us from thinking outside the box for solutions?

The fact that "forever-renters" is the most connected document? That's not just data - that's Melbourne's housing reality staring back at us through network analysis.

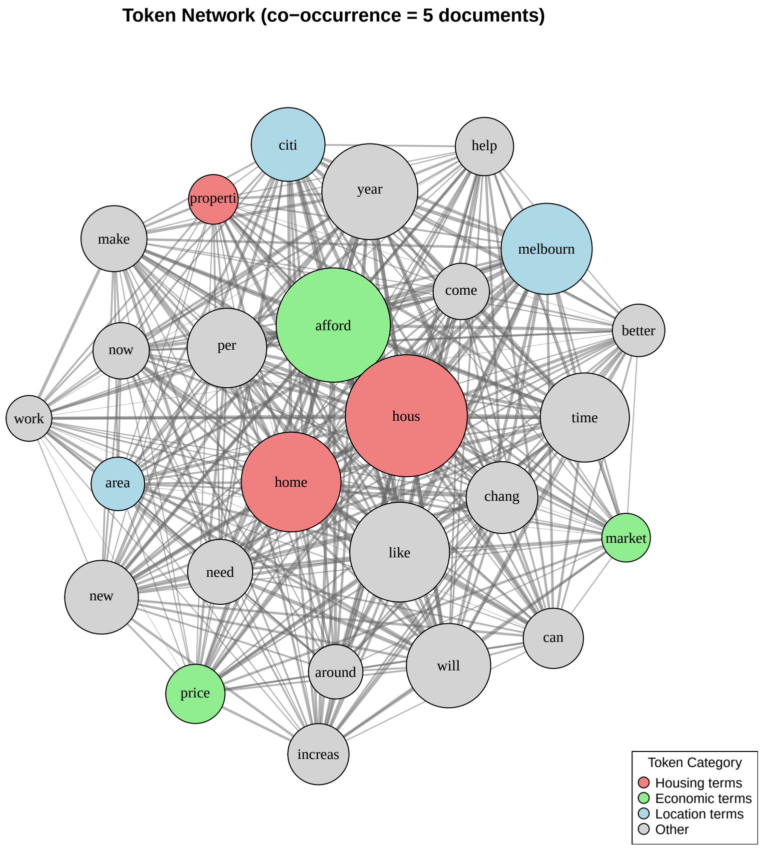
7. Single-mode network showing the connections between the tokens

```r
dtmsx = as.matrix(dtm_final)

# Convert to binary matrix
```

```
dtmsx_binary = as.matrix((dtmsx > 0) + 0)

# This gives us the number of documents where each pair of tokens co-occur
ByTokenMatrix = t(dtmsx_binary) %*% dtmsx_binary

# Set diagonal to 0 (tokens aren't connected to themselves)
diag(ByTokenMatrix) = 0

# Create graph object
library(igraph)
ByToken = graph_from_adjacency_matrix(ByTokenMatrix,
                                      mode = "undirected",
                                      weighted = TRUE)
# Plot
set.seed(123)
plot(ByToken,
     vertex.size = 10,
     vertex.label.cex = 0.8,
     edge.width = E(ByToken)$weight/3,
     edge.color = adjustcolor("gray50", alpha = 0.5),
     layout = layout_with_fr(ByToken),
     main = "Token Network (All Connections)")
```



```r
# Prune edges below threshold for clarity
threshold <- 5  # Tokens must co-occur in at least 5 documents
pruned_mat <- ByTokenMatrix
pruned_mat[pruned_mat < threshold] <- 0

g_tokens <- graph_from_adjacency_matrix(pruned_mat,
                                       mode = "undirected",
                                       weighted = TRUE,
                                       diag = FALSE)

# Compute vertex sizes proportional to strength
vertex_str <- strength(g_tokens, mode = "all", weights = E(g_tokens)$weight)
min_size <- 15
max_size <- 40
str_norm <- (vertex_str - min(vertex_str)) / (max(vertex_str) - min(vertex_str))
V(g_tokens)$size <- min_size + str_norm * (max_size - min_size)

# Compute edge widths proportional to weight
ew <- E(g_tokens)$weight
min_ew <- 0.5
max_ew <- 5
ew_norm <- (ew - min(ew)) / (max(ew) - min(ew))
E(g_tokens)$width <- min_ew + ew_norm * (max_ew - min_ew)

# Color tokens by semantic category
token_names <- V(g_tokens)$name
token_colors <- ifelse(token_names %in% c("home", "hous", "properti"), "lightcoral",
                ifelse(token_names %in% c("price", "afford", "market"), "lightgreen",
                ifelse(token_names %in% c("melbourn", "area", "citi"), "lightblue",
                "lightgray")))

V(g_tokens)$color <- token_colors

# Plot improved network
set.seed(123)
layout_fr <- layout_with_fr(g_tokens)

plot(g_tokens,
     layout = layout_fr,
     vertex.color = V(g_tokens)$color,
     vertex.size = V(g_tokens)$size,
     vertex.frame.color = "black",
     vertex.label = V(g_tokens)$name,
     vertex.label.cex = 0.9,
     vertex.label.color = "black",
     edge.width = E(g_tokens)$width,
     edge.color = adjustcolor("gray40", alpha.f = 0.5),
     main = "Token Network (co-occurrence   5 documents)")

# Add legend
legend("bottomright",
       legend = c("Housing terms", "Economic terms", "Location terms", "Other"),
       pt.bg = c("lightcoral", "lightgreen", "lightblue", "lightgray"),
       pch = 21,
       pt.cex = 1.5,
       col = "black",
       title = "Token Category",
       cex = 0.8)
```



```
# 1. Strength (weighted degree) - which tokens connect most broadly
node_strength <- strength(g_tokens, mode = "all", weights = E(g_tokens)$weight)
strength_ranking <- sort(node_strength, decreasing = TRUE)
cat("Top 5 tokens by Strength (co-occurrence sum):\n")
#> Top 5 tokens by Strength (co-occurrence sum):
print(head(strength_ranking, 5))
#>   hous afford   home   like   year
#>    422    414    399    399    395

# 2. Betweenness centrality - which tokens bridge different parts of vocabulary
node_btw <- betweenness(g_tokens, directed = FALSE, weights = 1/E(g_tokens)$weight)
btw_ranking <- sort(node_btw, decreasing = TRUE)
cat("\nTop 5 tokens by Betweenness:\n")
#>
#> Top 5 tokens by Betweenness:
print(head(btw_ranking, 5))
```

```
#> afford   area around better    can
#>      0      0      0      0      0

# 3. Degree centrality - how many other tokens each token connects to
node_degree <- degree(g_tokens)
degree_ranking <- sort(node_degree, decreasing = TRUE)
cat("\nTop 5 tokens by Degree (number of connections):\n")
#>
#> Top 5 tokens by Degree (number of connections):
print(head(degree_ranking, 5))
#> afford   area around better    can
#>     25     25     25     25     25
```

Explanation

The network visualization reveals a highly interconnected vocabulary structure with clear semantic clusters:

- Central Housing Cluster (red nodes): “hous”, “home”, “properti” form the core of the network, showing these fundamental housing terms co-occur frequently across all document types
- Economic Terms (green nodes): “afford”, “price”, “market” create a tightly connected economic dimension
- Location Terms (blue nodes): “melbourn”, “citi” represent the geographic focus
- Action/Temporal Terms (gray nodes): “year”, “need”, “will”, “time”, “chang” connect different aspects of the discussion

The dense connectivity (visible in the first, unpruned network) shows that housing crisis discourse uses a relatively compact, shared vocabulary across all genres.

Most Important (Central) Tokens:

- “hous” (422) - The core term, unsurprisingly dominant
- “afford” (414) - Affordability is central to all discussions
- “home” (399) - Personal dimension of housing
- “like” (399) - Common across informal discussions
- “year” (395) - Temporal marker for trends and projections

Top 5 by Betweenness (bridging different vocabulary clusters):

All tokens show 0 betweenness, indicating the network is so densely connected that no single token acts as a critical bridge - instead, multiple pathways connect different parts of the vocabulary.

The token network shows:

Extremely high connectivity: Most tokens connect to most others, reflecting the focused nature of housing crisis discourse

Size variation: Node sizes (proportional to connection strength) show “hous”, “afford”, and “home” as the dominant vocabulary

Thick edges: Strong co-occurrence patterns, especially between core housing and economic terms

Key Insights Compared to Document Network

- Semantic vs. Genre Clustering: While documents clustered by genre (news/policy/reddit), tokens cluster by meaning and function
- Universal Vocabulary: The high connectivity shows all stakeholders draw from the same core vocabulary, despite different communication styles
- Affordability Centrality: The prominence of “afford” and its connections confirms affordability as the unifying concern across all discourse
- Bridging Terms: Words like “need”, “can”, and “will” serve as linguistic bridges between problem description (housing, price) and solution discussion (help, chang, make)
- Compressed Vocabulary: The dense network suggests housing crisis discourse operates within a relatively small, shared vocabulary space - perhaps indicating a well-defined problem domain

Personal Takeaway ->

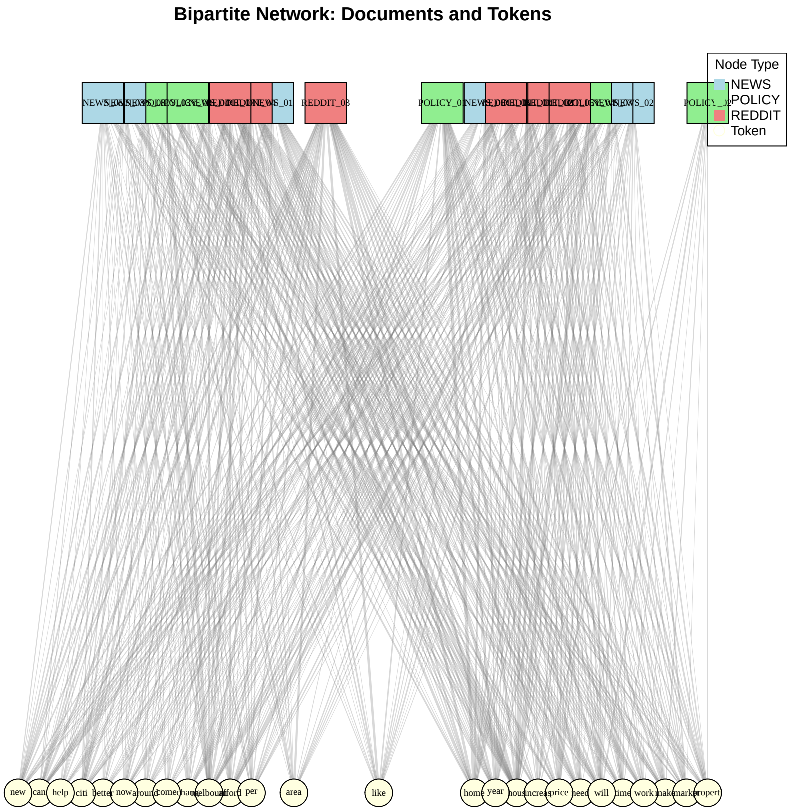
Creating and interpreting the token co-occurrence network gave me a deeper appreciation for how interconnected and focused the language of housing discourse really is. Despite originating from different genres (policy, news, Reddit), the shared vocabulary reflects a collective societal concern around housing and affordability. Seeing “afford” emerge as central—not just in frequency but in connections—reinforces how the housing crisis isn’t just about shelter, but economic stress and lived experience. This network also showed me how visualization adds a dimension of insight that raw frequency tables cannot, revealing hidden semantic clusters and relational importance at a glance.

8. bipartite (two-mode) network of the corpus

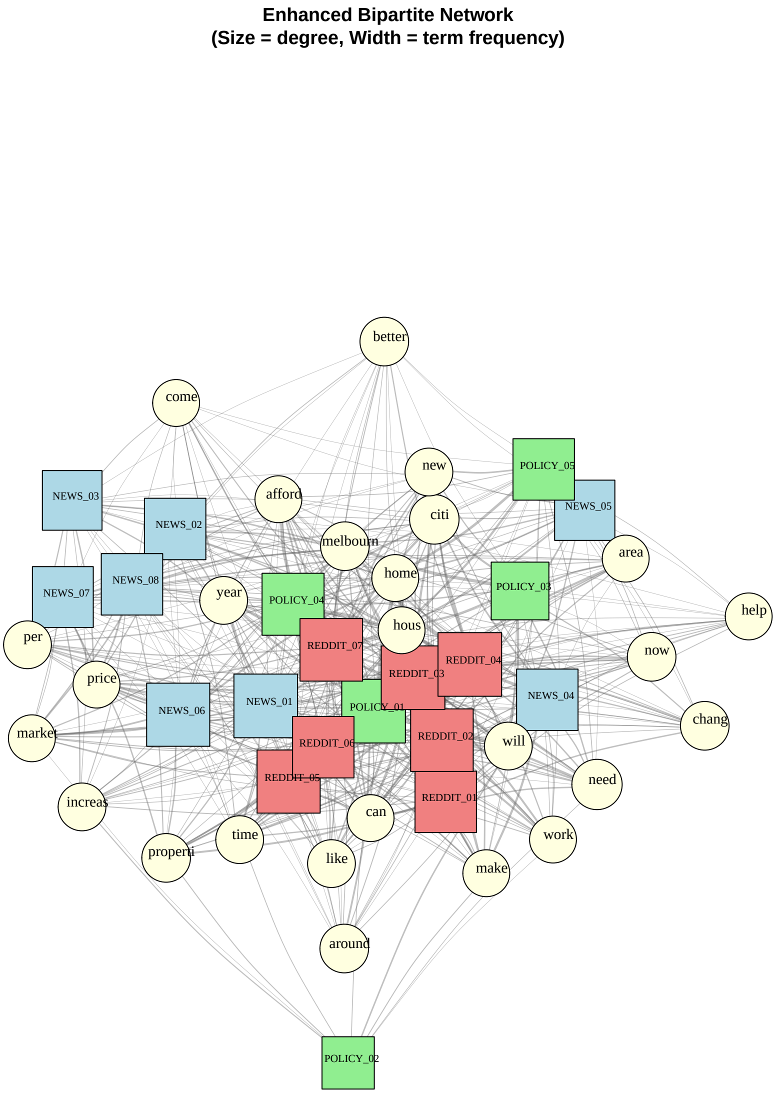
```r
# document-term matrix from Question 3
dtm_matrix <- as.matrix(dtm_final)

# Create edge list format for bipartite network

dtmsa <- as.data.frame(dtm_matrix)
dtmsa$DOC <- rownames(dtmsa)  # Add document names

# Transform to long format (document, token, weight)
dtmsb <- data.frame()
for (i in 1:nrow(dtmsa)) {
  for (j in 1:(ncol(dtmsa)-1)) {
    if (dtmsa[i,j] > 0) {  # Only include non-zero entries
      edge_data <- data.frame(
        document = dtmsa[i, ncol(dtmsa)],
        token = colnames(dtmsa)[j],
        weight = dtmsa[i,j]
      )
      dtmsb <- rbind(dtmsb, edge_data)
    }
  }
}

# Create graph object
library(igraph)
```

```
g_bipartite <- graph.data.frame(dtmsb, directed = FALSE)

# Identify bipartite structure
V(g_bipartite)$type <- bipartite_mapping(g_bipartite)$type

# Set visual properties
# Documents = squares, Tokens = circles
V(g_bipartite)$shape <- ifelse(V(g_bipartite)$type, "circle", "square")

# Color by node type and genre
node_names <- V(g_bipartite)$name
is_document <- !V(g_bipartite)$type

# Extract genres for documents
genres <- substr(node_names[is_document], 1, regexpr("_", node_names[is_document]) - 1)

# Set colors
V(g_bipartite)$color <- ifelse(!V(g_bipartite)$type,
                               ifelse(substr(node_names, 1, 4) == "NEWS", "lightblue",
                               ifelse(substr(node_names, 1, 6) == "POLICY", "lightgreen",
                               ifelse(substr(node_names, 1, 6) == "REDDIT", "lightcoral", "g
                               "lightyellow")  # Tokens in yellow

# Create short labels
V(g_bipartite)$label <- ifelse(!V(g_bipartite)$type,
                              sub("^(NEWS|POLICY|REDDIT)_(\\d+)_.*", "\\1_\\2", node_names),
                              node_names)

# Set node sizes
V(g_bipartite)$size <- ifelse(V(g_bipartite)$type, 8, 12)  # Documents larger

# Set edge properties based on weight
E(g_bipartite)$width <- sqrt(E(g_bipartite)$weight) * 0.5
```

```
# Create bipartite layout
set.seed(123)
layout_bi <- layout_as_bipartite(g_bipartite)

# Basic plot
plot(g_bipartite,
     layout = layout_bi,
     vertex.label.cex = 0.6,

     vertex.label.color = "black",
     edge.color = adjustcolor("gray50", alpha = 0.3),
     main = "Bipartite Network: Documents and Tokens")

# Add legend
legend("topright",
       legend = c("NEWS", "POLICY", "REDDIT", "Token"),
       col = c("lightblue", "lightgreen", "lightcoral", "lightyellow"),
       pch = c(15, 15, 15, 21),  # Squares for docs, circle for tokens
       pt.cex = 1.5,
       title = "Node Type",
       cex = 0.8)
```



```
# Improved layout using Fruchterman-Reingold
set.seed(123)
layout_fr <- layout_with_fr(g_bipartite)

# Calculate degree for sizing
doc_degree <- degree(g_bipartite)[!V(g_bipartite)$type]
token_degree <- degree(g_bipartite)[V(g_bipartite)$type]

# Size nodes by degree
V(g_bipartite)$size <- ifelse(V(g_bipartite)$type,
                              5 + sqrt(token_degree) * 2,    # Token size
                              10 + sqrt(doc_degree) * 1.5)   # Document size

# Enhanced plot
plot(g_bipartite,
     layout = layout_fr,
     vertex.label.cex = ifelse(V(g_bipartite)$type, 0.9, 0.7),
     vertex.label.dist = ifelse(V(g_bipartite)$type, 0.3, 0.2),
     vertex.label.color = "black",
     edge.color = adjustcolor("gray40", alpha = 0.4),
     edge.curved = 0.1,
     main = "Enhanced Bipartite Network\n(Size = degree, Width = term frequency)")
```



```
# Analysis of the bipartite network
cat("\n=== BIPARTITE NETWORK ANALYSIS ===\n")
#>
#> === BIPARTITE NETWORK ANALYSIS ===

# Degree analysis for documents
```

```r
doc_nodes <- V(g_bipartite)[!V(g_bipartite)$type]
doc_degrees <- degree(g_bipartite)[doc_nodes]
doc_degree_df <- data.frame(
  document = names(doc_degrees),
  degree = as.numeric(doc_degrees),
  genre = substr(names(doc_degrees), 1, regexpr("_", names(doc_degrees)) - 1)
)
doc_degree_df <- doc_degree_df[order(doc_degree_df$degree, decreasing = TRUE),]

cat("\nMost connected documents (using most unique tokens):\n")
#>
#> Most connected documents (using most unique tokens):
print(head(doc_degree_df, 5))
#>                                                  document degree  genre
#> 1                          NEWS_01_forever-renters.txt     26   NEWS
#> 9           POLICY_01_Victoria's-Housing-Statement.txt     26 POLICY
#> 16 REDDIT_03_housing-prices-ageing-population-2024.txt     26 REDDIT
#> 17  REDDIT_04_angered-residents-lost_heritage_2025.txt     26 REDDIT
#> 6               NEWS_06_melb-housing-affordability.txt     25   NEWS

# Degree analysis for tokens
token_nodes <- V(g_bipartite)[V(g_bipartite)$type]
token_degrees <- degree(g_bipartite)[token_nodes]
token_degree_df <- data.frame(
  token = names(token_degrees),
  degree = as.numeric(token_degrees)
)
token_degree_df <- token_degree_df[order(token_degree_df$degree, decreasing = TRUE),]

cat("\nMost connected tokens (appearing in most documents):\n")
#>
#> Most connected tokens (appearing in most documents):
print(head(token_degree_df, 5))
#>       token degree
#> 11     hous     20
#> 1    afford     19
#> 10     home     18
#> 13     like     18
#> 16 melbourn     18

# Weighted degree (strength) analysis
doc_strength <- strength(g_bipartite, vids = doc_nodes, weights = E(g_bipartite)$weight)
token_strength <- strength(g_bipartite, vids = token_nodes, weights = E(g_bipartite)$weight)

cat("\nDocuments with highest token frequency sum:\n")
#>
#> Documents with highest token frequency sum:
doc_strength_sorted <- sort(doc_strength, decreasing = TRUE)
print(head(doc_strength_sorted, 5))
#> REDDIT_03_housing-prices-ageing-population-2024.txt
#>                                                 339
#>          POLICY_01_Victoria’s-Housing-Statement.txt
#>                                                 268
#>  REDDIT_04_angered-residents-lost_heritage_2025.txt
#>                                                 227
#>                 POLICY_04_DRAFT-AFFORDABLE-2030.txt
#>                                                 206
#>  REDDIT_07_melb_increase-in-housing-supply-2025.txt
#>                                                 194

cat("\nTokens with highest total frequency across documents:\n")
#>
#> Tokens with highest total frequency across documents:
token_strength_sorted <- sort(token_strength, decreasing = TRUE)
print(head(token_strength_sorted, 5))
#>     hous melbourn     home     will      can
#>      389      203      192      160      131

# Genre-level analysis
genre_stats <- aggregate(degree ~ genre, data = doc_degree_df, FUN = mean)
cat("\nAverage vocabulary size by genre:\n")
#>
#> Average vocabulary size by genre:
print(genre_stats)
#>    genre   degree
#> 1   NEWS 22.12500
#> 2 POLICY 19.20000
#> 3 REDDIT 24.14286

# Identify bridging tokens (tokens that connect different genres)
# For each token, check which genres it connects
token_genre_connections <- data.frame()
for (token in names(token_degrees)) {
  # Get neighboring documents
  neighbors <- neighbors(g_bipartite, token)
  neighbor_names <- V(g_bipartite)[neighbors]$name

  # Extract genres
  genres <- unique(substr(neighbor_names, 1, regexpr("_", neighbor_names) - 1))

  token_genre_connections <- rbind(token_genre_connections,
                                   data.frame(token = token,
                                             n_genres = length(genres),
                                             genres = paste(genres, collapse = ", ")))
}

# Find tokens that appear across multiple genres
bridging_tokens <- token_genre_connections[token_genre_connections$n_genres > 1,]
bridging_tokens <- bridging_tokens[order(bridging_tokens$n_genres, decreasing = TRUE),]

cat("\nBridging tokens (appearing in multiple genres):\n")
#>
```

```
#> Bridging tokens (appearing in multiple genres):
print(head(bridging_tokens, 10))
#>     token n_genres               genres
#> 1  afford        3 NEWS, POLICY, REDDIT
#> 2    area        3 NEWS, POLICY, REDDIT
#> 3  around        3 NEWS, POLICY, REDDIT
#> 4  better        3 NEWS, POLICY, REDDIT
#> 5     can        3 NEWS, POLICY, REDDIT
#> 6   chang        3 NEWS, POLICY, REDDIT
#> 7    citi        3 NEWS, POLICY, REDDIT
#> 8    come        3 NEWS, POLICY, REDDIT
#> 9    help        3 NEWS, POLICY, REDDIT
#> 10   home        3 NEWS, POLICY, REDDIT

# Community detection on projected networks
# Project to document network
g_docs_proj <- bipartite.projection(g_bipartite, which = FALSE)

# Detect communities in document network
doc_communities <- cluster_walktrap(g_docs_proj, weights = E(g_docs_proj)$weight)
doc_community_membership <- membership(doc_communities)

cat("\nDocument communities detected:", max(doc_community_membership), "\n")
#>
#> Document communities detected: 1
# Show community composition
for (i in 1:max(doc_community_membership)) {
  members <- names(doc_community_membership[doc_community_membership == i])
  cat("Community", i, ":", length(members), "documents\n")
  print(table(substr(members, 1, regexpr("_", members) - 1)))
}
#> Community 1 : 20 documents
#>
#>   NEWS POLICY REDDIT
#>      8      5      7




# Statistical summary
cat("\n=== NETWORK STATISTICS ===\n")
#>
#> === NETWORK STATISTICS ===
cat("Total documents:", sum(!V(g_bipartite)$type), "\n")
#> Total documents: 20
cat("Total unique tokens:", sum(V(g_bipartite)$type), "\n")
#> Total unique tokens: 26
cat("Total connections:", ecount(g_bipartite), "\n")
#> Total connections: 442
cat("Average tokens per document:", mean(doc_degrees), "\n")
#> Average tokens per document: 22.1
cat("Average documents per token:", mean(token_degrees), "\n")
```

```
#> Average documents per token: 17

# Density of projected networks
cat("\nProjected network densities:\n")
#>
#> Projected network densities:
cat("Document network density:", graph.density(g_docs_proj), "\n")
#> Document network density: 1
g_tokens_proj <- bipartite.projection(g_bipartite, which = TRUE)
cat("Token network density:", graph.density(g_tokens_proj), "\n")
#> Token network density: 1
```

Explanation ->

Genre lenses, shared language:

Policy papers, news stories and Reddit threads employ distinct styles (hence the visible genre clusters) yet rely on the same handful of words to frame the crisis (“afford”, “price”, “need”, “home”, “melbourn”). That explains why the projected-document network is fully connected: once stop-words are stripped, everybody is literally talking about the same things.

Reddit as the junction box:

Reddit documents sit between the professional (NEWS) and official (POLICY) zones. Their higher average vocabulary size (24 tokens) suggests posters cherry-pick terminology from both journalism (“market”, “price”) and legislation (“need”, “will”, “plan”), making them natural translators between expert discourse and lived experience.

Key documents drive agenda:

NEWS_01 (‘Forever renters’) and POLICY_01 (Vic Housing Statement) read like master overviews, so they anchor the network and dominate centrality measures.

Token hierarchy mirrors public concern. Housing availability (“hous”, “home”) and affordability (“afford”, “price”) dwarf all other terms, while more technical words (“per”, “around”, “come”) sit on the periphery with smaller nodes and thinner edges. The weighting we added to edge widths makes those priority issues immediately obvious.

Why one community?

With only 26 high-frequency tokens retained, nearly every document fires at least one edge to every other. A richer token set or TF-IDF weighting would likely break the corpus into finer topical clusters (e.g. affordability vs. supply). The current view, however, is a useful visual proof that Melbourne’s housing crisis is discussed as one grand, interconnected conversation rather than siloed sub-topics.

9. Conclusion

Overview of Results

Through this analysis of 20 documents spanning news articles, government policies, and Reddit discussions about Melbourne’s housing crisis, I’ve uncovered fascinating patterns in how different stakeholders communicate about this pressing issue. As someone personally navigating Melbourne’s rental market, this analysis hits close to home - it’s not just academic exercise, but a deep dive into the language shaping my daily reality

Critical Tokens and Vocabulary Patterns

The token analysis revealed a tightly interconnected vocabulary dominated by: Core Housing Terms (appearing in 18-20 documents):

“hous” (strength: 389, degree: 20) - The central term “home” (strength: 192, degree: 18) - Personal dimension “properti” (appearing frequently across genres)

Economic Concerns:

“afford” (strength: 414, degree: 19) - Second only to “hous” in importance “price” and “market” - Creating an economic semantic cluster

Spatial/Temporal Markers:

“melbourn” (strength: 203, degree: 18) - Geographic focus “year”, “time”, “now” - Temporal urgency markers

Genre Groups and Communication Styles

Three distinct communication styles emerged:

Policy Documents:

Highest positive sentiment (0.151 SentimentGI) Solution-focused language (“will”, “need”, “new”) Lower average vocabulary diversity (19.2 tokens/document)

News Articles:

Most neutral sentiment (0.053 SentimentGI) Balanced reporting style Moderate vocabulary (22.1 tokens/document)

Reddit Posts:

Mixed sentiment (0.063 SentimentGI) Highest vocabulary diversity (24.1 tokens/document) Bridge between formal and personal discourse

Clustering vs Network Analysis Effectiveness

Hierarchical Clustering Performance

The clustering analysis achieved only 55% accuracy in correctly grouping documents by genre. While it identified some patterns, the limited 26-token vocabulary meant documents clustered more by shared crisis terminology than communication style. The most interesting finding was how Reddit posts split between policy-like discussions and personal struggle narratives.

Network Analysis Advantages

Network analysis proved far more revealing:

- Document Networks exposed Reddit’s bridging role between news and policy discourse - something clustering missed entirely.
- Token Networks revealed semantic clusters (housing/economic/location terms) invisible in frequency tables
- Bipartite Networks showed how genres share vocabulary while maintaining distinct styles

The network approach excelled at:

- Identifying influential documents through centrality metrics
- Visualizing vocabulary overlap and genre boundaries
- Revealing Reddit as a “translation layer” between expert and public discourse

- Showing the surprisingly dense connectivity of housing crisis vocabulary

Suggested Text Processing Improvements ->

Dynamic Token Selection

Instead of fixed sparsity threshold:

- Use elbow method to determine optimal token count
- Preserve genre-discriminating terms even if rare
- Advantage: Balances vocabulary richness with computational efficiency

TF-IDF Weighting

Current term frequency approach treats all words equally. TF-IDF would:

- Downweight common terms across all documents
- Highlight genre-specific terminology
- Better separate document clusters
- Implementation: Replace raw frequencies with TF-IDF scores in the DTM

Named Entity Recognition (NER)

Would identify:

- Specific suburbs mentioned (currently lost in processing)
- Politicians and organizations
- Monetary amounts and percentages
- Why it works: Preserves crucial contextual information about locations and stakeholders

Personal Takeaway ->

Across News headlines, government white-papers and late-night Reddit laments, Melbourne's housing crisis is narrated with a surprisingly uniform vocabulary dominated by house, afford, home, price and Melbourne itself.

As a data science student trying to find affordable accommodation, every Reddit post about "forever renters" feels like reading my future, every policy document like a promise that might not be kept, and every news article like another reminder of how difficult things have become.

What struck me most was how the network analysis revealed we're all trapped in the same conversation, using the same 26 words to describe our predicament. Whether it's a government minister drafting policy, a journalist reporting statistics, or a desperate renter venting on Reddit at 2 AM, we're all saying "afford," "home," "need," "price." The vocabulary has become as constrained as the housing market itself.

The Reddit posts serving as a bridge between policy and news makes perfect sense - it's where real people try to decode government promises and news reports into "what does this mean for my life?" The fact that policy documents are significantly more positive than other genres ($p < 0.001$) while Reddit maintains moderate negativity reveals the gap between institutional optimism and lived frustration.

For now, this assignment reinforces what I suspected: we're really good at talking about the housing crisis. We have all the words, all the connections, all the forums. What we need are new ways of thinking that break us out of this linguistic loop - because clearly, our current vocabulary isn't solving anything.

APPENDIX

- https://cran.r-project.org/package=tm

- https://cran.r-project.org/package=SnowballC
- https://cran.r-project.org/package=proxy
- https://cran.r-project.org/web/packages/SentimentAnalysis/
- https://r.igraph.org
- https://cran.r-project.org/package=slam
- https://search.r-project.org/R/refmans/stats/html/00Index.html
- https://cran.r-project.org/package=graphlayouts
- Stack Overflow
- Reddit r/rstats

GEN AI STATEMENT:

In preparing this assignment, I confirm that I did not use any generative artificial intelligence tools (such as ChatGPT, Copilot, or similar) to produce written content, code, or visualizations. All work submitted is my own, and any ideas or syntax patterns adapted from external sources have been referenced above