Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

NLP 💬

[Dr. Hanjo Odendaal]

Photo by Stefan Els

Setup

First things first, lets install the necessary packages that we will need in this lesson as well as create the necessary files to fit everything in a `targets` pipeline:

Step 1: Install Packages

```
install.packages("tidytext", Ncpus = 4)
install.packages("quanteda", Ncpus = 4)
install.packages("topicmodels", Ncpus = 4)
install.packages("ggwordcloud", Ncpus = 4)
```

Step 2: Create Folder for Analysis

```
Texevier::create_template(directory = "~/Downloads"
                          template_name = "NLP")
```

Step 3: Create `Readme.qmd` file

```
---
title: "My readme"
format: html
editor_options:
  chunk_output_type: console
execute:
  echo: false
  eval: false
---
```

```
library(tidyverse)
library(targets)
library(dbbasic)
library(tidytext)
library(quanteda)
library(topicmodels)
library(ggwordcloud)
```

Setup

Step 4: Create .Renviron file

`usethis::edit_r_environ()`

```
gp_data = datascience
gp_user = datascience
gp_pass = f5VPEC8nsU01QKbSxSfv
gp_host = localhost
gp_port = 3000
```

Step 6: Restart Rstudio and Test!

```
db_query(
  "SELECT * FROM
   pg_catalog.pg_tables
  WHERE schemaname ='public';"
  , db = "psql_datascience")
```

Phew

NLP Basics

In this session we will be exploring the basics of Natural Language Processing:

Exploratory Analysis:

Concordancing

- Extraction of words from a given text or texts conditional on a context window

Ngrams

- Basic word counts from texts
- Creating word clouds from the text

Modeling:

Sentiment Analysis

- Using dictionary methods

Topic Modeling (Bonus)

- Text > Tokens > Document Frequency Matrix > Topic Model

First things first: Data

Our data is currently stored in a cloud Database. Pull the **sales** data into your session using the `dbbasic` package. Remember to open a `dev.sql` file to test your queries from inside R!:

- Quick look at what the database table contains again:

```sql
-- !preview conn=db_connect(db = "psql_datascience")

SELECT * FROM gumtree_clean LIMIT 10;
```

type	ad_id	ad_url	location	for_sale_by	dwelling_type	bedrooms	bathrooms	size_sqm	parking	price	available_from	for_rent_by	furnished	smoking	pet_friendly	description
sales	1b4d044419256d4999586eaee6baca8c	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	3	3	546	garage	39995000	NA	NA	NA	NA	NA	This incredible property in a reno
sales	991f741a8f3619089ba10cae7d5b4072	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	1	2	211	garage	11500000	NA	NA	NA	NA	NA	RE/MAX Living operates in terms
sales	f9a57d1d6310bc561da6d86ec3fb6f53	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	2	2	118	covered	16995000	NA	NA	NA	NA	NA	Asking Price: R 16 995 000Nestle
sales	8b3c83c89488ee15eb5a441360ef44df	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	house	3	3	589	garage	19500000	NA	NA	NA	NA	NA	RE/MAX Living operates in terms
sales	0a0f80664100990b7ebc40169081fd79	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	house	1	1	138	covered	14500000	NA	NA	NA	NA	NA	RE/MAX Living operates in terms
sales	812555990c3cede1ccf24b07fc85800a	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	2	2	118	garage	14600000	NA	NA	NA	NA	NA	RE/MAX Living operates in terms
sales	c2c7c97fbd2cc996cee5b0821177a259	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	4	3	301	covered	12995000	NA	NA	NA	NA	NA	Exclusive Mandate | Asking Price:
sales	cc71a7903344e3d7debca37f13a48d01	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	2	2	118	covered	16995000	NA	NA	NA	NA	NA	Aurum Luxury Residences introdu
sales	844b53eb84712852ce4241ddca268f2c	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	1	2	69	NA	4650000	NA	NA	NA	NA	NA	RE/MAX Living operates in terms
sales	2dd9514800082e2e3ce91f973151a338	https://www.gumtree.co.za/a-houses-flats-for-sale/b...	bantry_bay_atlantic_seaboard	agency	apartment	1	1	NA	NA	4650000	NA	NA	NA	NA	NA	Nestled along the picturesque Be

First things first: Data

- What are the different variables within the `type` column?

```
SELECT
    type,
    COUNT(*)
FROM gumtree_clean
GROUP BY
    type
LIMIT 10;
```

- Filter the dataset to only bring in **sales** into R environment:

```
gumtree <- db_query(" SELECT * FROM gumtree_clean WHERE type = 'sales'",
                    db = "psql_datascience")

gumtree_texts <- gumtree %>% select(ad_id, description) %>%
  mutate(description = tolower(description))
```

Case Study: 'Ocean'

Concordancing

As mentioned, "concordancing" is where we would like to see what how words a are used in context, or "keywords-in-contexts" analysis. Let us use the `kwic` function from `quanteda` to see how the word **ocean** is used in the text.

```
ocean_kwic <- kwic(
  # define text
  gumtree_texts$description,
  # define search pattern
  pattern = "ocean",
  # define context window size
  window = 5) %>%
  as_tibble

# [main]>ocean_kwic
# A tibble: 1,772 × 7
#    docname  from    to pre                               keyword post                          pattern
#    <chr>   <int> <int> <chr>                             <chr>   <chr>                         <fct>
#  1 text1      16    16 has views overlooking the open    ocean   , robben island and on        ocean
#  2 text3     116   116 stunning vistas of the azure      ocean   that greet you from every     ocean
#  3 text5     134   134 unobstructed views of the atlantic ocean   from every angle along with   ocean
#  4 text7      85    85 panoramic views of the atlantic   ocean   and lion's head , emphasising ocean
#  5 text7     281   281 views of the majestic atlantic    ocean   . designed with entertainment in ocean
# # ℹ 1,762 more rows
# # ℹ Use `print(n = ...)` to see more rows
```

Concordancing

We can also extract exact phrases, by using the function `phrase()`:

```
ocean_kwic ← kwic(
  # define text
  gumtree_texts$description,
  # define search pattern
  pattern = phrase("blue ocean"),
  # define context window size
  window = 5) %>%
  as_tibble
```

```
# [main]>ocean_kwic
# A tibble: 6 × 7
#   docname    from    to pre                                  keyword    post                                pattern
#   <chr>     <int> <int> <chr>                                <chr>      <chr>                               <fct>
# 1 text5849     78    79 the view of the wide                 blue ocean stretching to the horizon .         blue ocean
# 2 text6106    124   125 with sliding windows overlooking the blue ocean ! with its stunning ocean           blue ocean
# 3 text6160    100   101 up to the most bright                blue ocean views . neutral colour palette      blue ocean
# 4 text12564    67    68 by the beauty of the                 blue ocean shimmering through glass windows and blue ocean
# 5 text12605   255   256 onto table mountain with the         blue ocean at its feet , complete              blue ocean
# 6 text14439    42    43 sands beach with its beautiful       blue ocean right on your doorstep ?            blue ocean
```

Concordancing

Exercise time!

- How many instances are there of "swimming pool"?

05:00

Ngrams

Ngrams forms the basis of most text analysis. It is the fundamentals of *tokenization* or breaking up texts into words or sequences of words. Lets take our ocean example a little bit further and analyse the top words (after *stopwords*) within the contexts of the 'ocean' in property ads.

We are going to use a really nice function from `tidytext` called `unnest_tokens` for this.

```
ocean_kwic ← kwic(gumtree_texts$description, pattern = "ocean",
                  window = 5) %>% as_tibble
```

- Combine the `pre` and `post` columns into one:

```
ocean_pre_post ← ocean_kwic %>% unite("text", c(pre, post)) %>% select(docname, text)
```

```
# A tibble: 1,772 × 2
#   docname text
#   <chr>   <chr>
# 1 text1   has views overlooking the open_, robben island and on
# 2 text3   stunning vistas of the azure_that greet you from every
# 3 text5   unobstructed views of the atlantic_from every angle along with
# 4 text7   panoramic views of the atlantic_and lion's head , emphasising
```

Ngrams

Now that we have our text ready, lets create `unigrams` or single word tokens from the text.

```
ocean_tokens <- ocean_pre_post %>% unnest_tokens(input = text, output = word, n = 1)
```

```
#  A tibble: 14,787 × 2
#    docname word
#    <chr>   <chr>
#  1 text1   has
#  2 text1   views
#  3 text1   overlooking
#  4 text1   the
```

⚠️ We need to now get rid of `stopwords` or words like 'the', 'a' or 'on' as these do not add contextualization.

```
ocean_tokens <- ocean_tokens %>%
  mutate(word = gsub("_", "", word)) %>% anti_join(stop_words, by = join_by(word))
```

```
# A tibble: 8,890 × 2
 #   docname word
 #   <chr>   <chr>
 # 1 text1   views
 # 2 text1   overlooking
 # 3 text1   open
 # 4 text1   robben
 # 5 text1   island
```

Plotting the word clouds

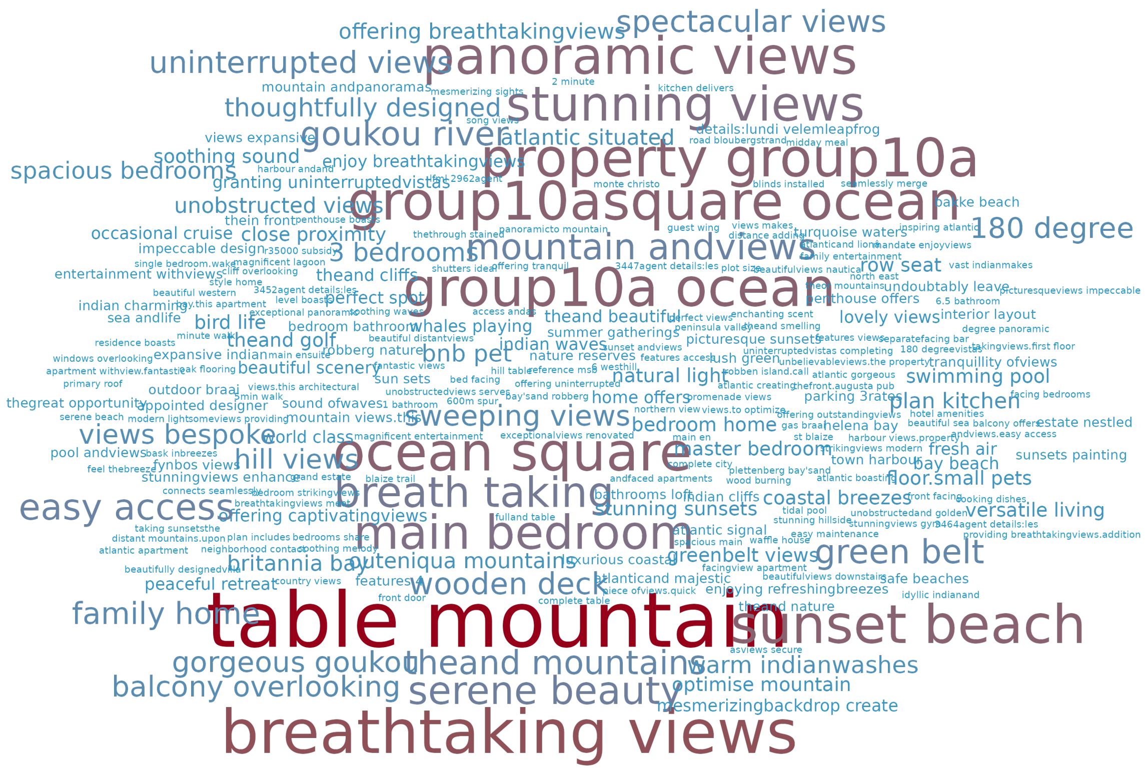
Stellenbosch University Iyunivesithi Universiteit

Once the text is in a nice tidy format, we can now do a lot with it... first lets plot the clouds to see what are the words closely associated around the 'ocean':

```
ocean_tokens %>%
  count(word, name = "obs", sort = TRUE) %>%
  sample_frac(weight = obs, size = 0.1) %>%
  ggplot(., aes(label = word, size = obs,
                color = obs)) +
  geom_text_wordcloud() +
  scale_color_gradient(low = "#189bcc",
                       high = "#960018") +
  scale_size_area(max_size = 20) +
  theme_minimal()
```

Plotting Bi-Grams (Bonus)

```r
ocean_pre_post %>%
  unnest_tokens(input = text, output = word,
                token = "ngrams", n = 2) %>%
  mutate(word = gsub("_", "", word)) %>%
  separate(word, c("word1", "word2"), sep = " ") %>
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  unite(bigram, word1, word2, sep = " ") %>%
  count(bigram, name = "obs", sort = TRUE) %>%
  sample_frac(weight = obs, size = 0.1) %>%
  ggplot(., aes(label = bigram, size = obs,
                color = obs)) +
  geom_text_wordcloud() +
  scale_color_gradient(low = "#189bcc",
                       high = "#960018") +
  scale_size_area(max_size = 20) +
  theme_minimal()
```



Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Modeling

Sentiment analysis in action

Provided that you now have the basics of text analysis and how to get from text > tokens, we can now apply some basic modeling techniques.

Sentiment analysis is a very nice kick-off point as it ranges in complexity from basic dictionary techniques (what we will be using) to intricate deep learning models. The following is an illustrative example of how sentiment is calculated using a news article and a dictionary approach:


Global capital flows into emerging markets including SA are drying up Andrew McNulty looks at how the changing game will affect investors For most of 2010 and 2009 SA and other emerging markets were popping champagne corks hot money was flowing in abundance most of last year global capital flowed strongly into emerging markets The strong global capital flows helped lift prices of these countries equities and other assets strengthened their currencies and underpinned growthBut the capital flows have dwindled or turned negative since last October Concerns about relative values rising inflation in these economies social unrest in North Africa and the Middle East and surging energy prices have turned investors against emerging markets That includes SA despite JSE equities resilience and the rands strength against a weak US dollar After the Reserve Banks monetary policy committee meeting last Thursday governor Gill Marcus said SA had experienced net sales of bonds and equities by non residents in the year to date totalling R192bn As the world economic recovery gains momentum and broadens investors views on the risks and opportunities have changedMarcus says the global recovery appears to have remained on track but the unresolved European debt crisis rising oil price and tragic events in Japan may moderate the pace of recovery in the near term Global inflation risks have increased particularly in emerging market economies some of which have raised shortterm interest rates The dwindling of foreign capital inflows has not led to steep falls in emerging equities but these markets are no longer rising and many have underperformed the developed markets over the past six monthsSince the end of September the MSCI emerging markets US index has risen 5 while the MSCI world index representing 24 developed markets gained 12 The JSE all share index rose by 7 in rand and by 10 in dollar termsSanlam group economist Jac Laubscher says capital will not necessarily keep flowing into emerging markets at a steady rate The inflows will depend on market circumstances he says I see the recent declines as a tactical or cyclical response to changes in relative values The positive longterm trend should continue However developed markets may outperform for a whileIn the recovery from the downturn investors had it easy They could buy emerging market equities or bonds and attain better returns than they could expect from developed markets where growth remains slow and risks are highMarkets such as the JSE were favourably rerated on recovering domestic growth improving company earnings and rising commodity prices Strengthening emerging market currencies enhanced returns for foreign investorsSome investors believe that the game has changed Despite the risks continued high unemployment and debt developed economies have returned to growth Company earnings have risen strongly and equity markets in developed economies may now offer better valueHow does the change in capital flows affect the growth outlook for emerging markets particularly SA and for investors The answers depend partly on how long the reversal of capital flows lasts and the impact of other developments such as the recent spike in the oil price Capital flowed into emerging markets after the financial crisis mainly because of their high growth and structural strength perceived lower risk and higher returns In many countries favourable economic fundamentals such as low fiscal deficits surpluses or small deficits in current accounts huge foreign exchange reserves and an expanding middle class should support strong longterm growthFor now the real questions are about their shortterm outlook and relative value which no longer looks as attractive against developed economies If capital outflows continue that could weaken the growth outlook and increase risks for some countries Capital flows to developing economies dipped in 2008 but have been in a strong upward trend since early in the past decade after the recovery from the 1997 1998 Asian crisisThe Institute of International Finance IIF says net private capital inflows into emerging markets rose from less than US200bn or below 15 of GDP in 2002 to 908bn or about 7 of GDP in 2010 An IIF report in January forecast the figure would rise to 960bn this year and to just over 1trillion in 2012It predicted slower growth in portfolio equity and bank inflows in this period as emerging equity markets weakened and capital controls and other credit measures curbed credit growth It identified two main risks Policymakers in emerging economies could make errors such as delaying or avoiding monetary or fiscal tightening to stall currency appreciation making a boom bust cycle more likely And developed economies could start raising rates sooner than expected reducing the relative attractiveness of emerging market yields That may be happening now The European Central Bank signalled earlier this month it may raise rates at its next meetingNonetheless the IIF forecasts continuing capital inflows mainly on the prospect of strong near term growth Developing nations particularly in Latin America and Asia have other attractions including increased political stability improved macroeconomic policies and greater structural strength than in the mid1990sA sudden sharp reversal of capital inflows was a feature of the Asian crisis almost 14 years ago International Monetary Fund IMF research shows net inflows to the Asian crisis countries were about 63 of their GDP in 1995 and 58 in 1996 In 1997 net outflows were 2 of GDP In 1998 they rose to 52 of GDPStructural weaknesses that grew over years of fast growth made several Asian countries including Korea Indonesia and Thailand particularly vulnerable It was partly a banking crisis Much of the capital inflows were in bank lending After a huge expansion of bank credit bad debts had ballooned Poor regulation of banking and financial sectors enabled banks to build liabilities in one country with assets in anotherThough the weaknesses were slightly different there was a common theme in Indonesia Korea and Thailand Alongside weak banking supervision they had a history of heavy government involvement in credit allocation These were reflected in poor lending practices and inflated asset pricesThere were also financial rigidities Some countries had large current account or trade deficits but maintained pegged exchange rates for too long Foreign lenders and investors contributed to asset bubbles as they failed to evaluate risks correctly After Thailands baht came under speculative pressure in the first half of 1997 in July that year the Thai government devalued the currency and announced a managed float The Thai crisis led to competitive currency weakness among other countries in the region forcing foreign lenders and investors to reassess risk The contagion spread across emerging markets and led to deep recessionsCould a similar pattern be repeated Asian economies have again experienced almost a decade of strong growth China and other East Asian countries have announced measures aimed at curbing the rise in property prices Many of the highgrowth developing countries now run current account surpluses or manageable deficits Many also hold huge foreign exchange reservesIn its October 2010 Global Financial Stability Report the IMF said rating agencies had downgraded developed country sovereign debt issues 25 times since early 2008 In 2010 there were upgrades of emerging market sovereign debt The IMF predicts the trend will continue A potential source of volatility is the carry trade which occurs when investors borrow money at low rates and invest it in assets in countries where rates are higher Australia New Zealand and SA have been favoured destinations for Japanese investors see story on page 32If rates start rising in Europe the US or Japan it could make the carry trade less attractive However countries such as Brazil India and Australia have raised shortterm rates many times in the past year SAs rates remain at 30year lows though at the bottom of the cycle SA is becoming less attractive as a yield play relative to some other emerging markets says Bureau for Economic Research BER senior economist Hugo PienaarThe BER said last month financial markets had been fairly aggressive in pricing in a local rate increase in SA before the end of this year The BER forecasts a first 50 basis point increase in November followed by more increases totalling 100 basis points in the first quarter of 2012 If foreign capital flows remain weak or negative it could force the Reserve Bank to raise rates earlier though it would depend on other factors such as the rand see story on page 37 A relatively strong rand acted as a buffer against rising energy and food prices Since the beginning of the year it has weakened by 8 against the euro but has been more stable against the weak dollar For now the local inflation outlook is benign The Banks monetary policy committee says inflation is expected to average 47 in 2011 and 57 in 2012However a combination of continued capital outflows high oil prices and a weaker rand if that occurs may change that view In its economic update earlier this year the IMF forecast emerging and developing economies would grow at 65 in 2011 and 2012 with developing Asia including China and India at 84 For advanced economies the forecast is 25 The BER forecasts SA will grow at 38 this year and 39 in 2012Many analysts believe the longterm outlook is compelling Global bank Standard Chartered chief economist Gerard Lyons says the world may have entered a new supercycle driven by industrialisation and urbanisation of emerging markets and global tradeHe says Chinas growth rate could average 69 over the next two decades to 2030 and India may grow faster Chinas income per capita will still be only half that of the US in 2030 leaving room for catchup Emerging markets now account for almost 40 of global consumption and more than twothirds of global growth However cycles can still be expected The IMF says if policymakers fail to head off nascent overheating pressures and asset price bubbles it would raise risks of a hard landing in these economies That would deal a serious blow to the global recovery Its little wonder that foreign investors have turned cautious about risk and value while these countries adjust to the risks of overheating and inflation

From the illustration, we can pick out that the article contains 93 negative words and 42 positive words:

$$\mathcal{A}_{it} = \frac{PositiveWords - NegativeWords}{PositiveWords + NegativeWords}$$

$$-0.37 = (42 - 93)/(42 + 93)$$

This news article is thus deemed to be mostly negative. Beware, there are directional biases within sentiment analysis and its a good idea to normalise the scores for analysis.

Sentiment analysis in action

Now that we have then basic idea of sentiment analysis using a dictionary approach, lets answer the following question:

- Are the ads for houses more positive than apartments?
- Is there a correlation between price and sentiment?

To conduct this analysis we have to perform the following steps:

1) Transform our text into tokens

2) Remove stopwords

3) Join in a sentiment dictionary

4) Group by advert type and analyse

Sentiment analysis: Houses vs Apartments

We start off by turning our text into tokens using `unnest_tokens`:

```
gumtree %>% select(dwelling_type, description) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words, by = "word")
```

Next, lets use the `bing` sentiment dictionary to determine the relative sentiment per advert:

```
tidytext::get_sentiments(c("bing", "afinn", "loughran", "nrc")[1])
# A tibble: 10 × 2
#   word         sentiment
#   <chr>        <chr>
# 1 trouble-free positive
# 2 displaced    negative
# 3 imposers     negative
# 4 achievible   positive
# 5 temptation   negative
# 6 cliche       negative
```

Sentiment analysis: Houses vs Apartments

Create the sentiment score per ad:

```
gumtree_sentiment ← gumtree %>%
  select(ad_id, dwelling_type, description) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words, by = "word") %>%
  left_join(get_sentiments("bing"), by = "word") %>%
  drop_na() %>%
  count(ad_id, dwelling_type, sentiment, name = "obs") %>%
  pivot_wider(names_from = "sentiment", values_from = "obs",
              values_fill = 0) %>%
  mutate(sentiment = (positive - negative)/(positive + negative))

# [main]>gumtree_sentiment
# A tibble: 14,562 × 5
#   ad_id                            dwelling_type positive negative sentiment
#   <chr>                            <chr>            <int>    <int>     <dbl>
# 1 00026b744459e17f11d5d66a9634f159 apartment            3        0     1
# 2 000513d19f7999faf08868e98d1a5dde house               12        0     1
# 3 0007b1cbdc9ef27ad3d663aaa5a11240 house                9        2     0.636
# 4 000a2cd6656b24763bd1fc416ea01b00 house                9        1     0.8
# 5 000b5a78eeb86987a0f8f4fba68fd568 house                3        1     0.5
```

Sentiment analysis: Houses vs Apartments

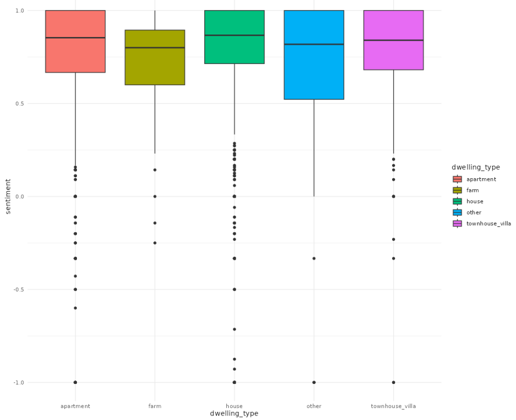
Using the sentiment data frame we are now able to test whether adverts for houses are more positive than apartments. To do this, we turn to stats 📊

- Boxplots and Density plots for visual
- Wilcoxon test for same continuous distribution (non-parametric version of a t-test)

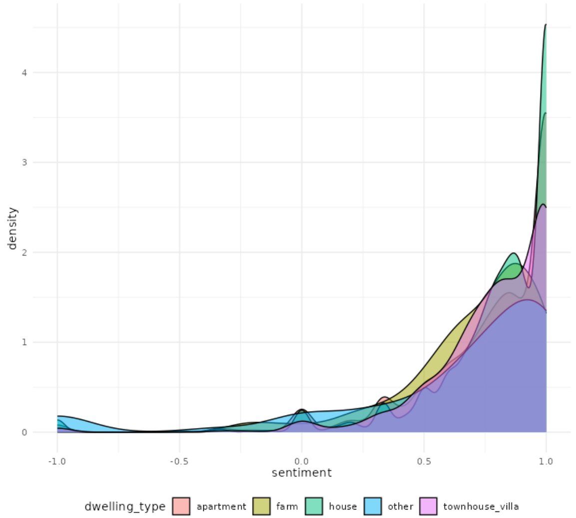
Sentiment analysis: Houses vs Apartments

Lets start with visual inspections:

```r
gumtree_sentiment %>%
  ggplot(., aes(dwelling_type, sentiment,
                fill = dwelling_type)) +
  geom_boxplot() +
  theme_minimal()
```




```r
gumtree_sentiment %>%
  ggplot(., aes(sentiment, fill = dwelling_type)) +
  geom_density(alpha = 0.5) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Sentiment analysis: Houses vs Apartments

It is good that we performed the visual inspection as one would have seen two things:

- The data is *not* normally distributed and as thus, we cannot use a *parametric* t-test
- There are other categories which we have to filter out

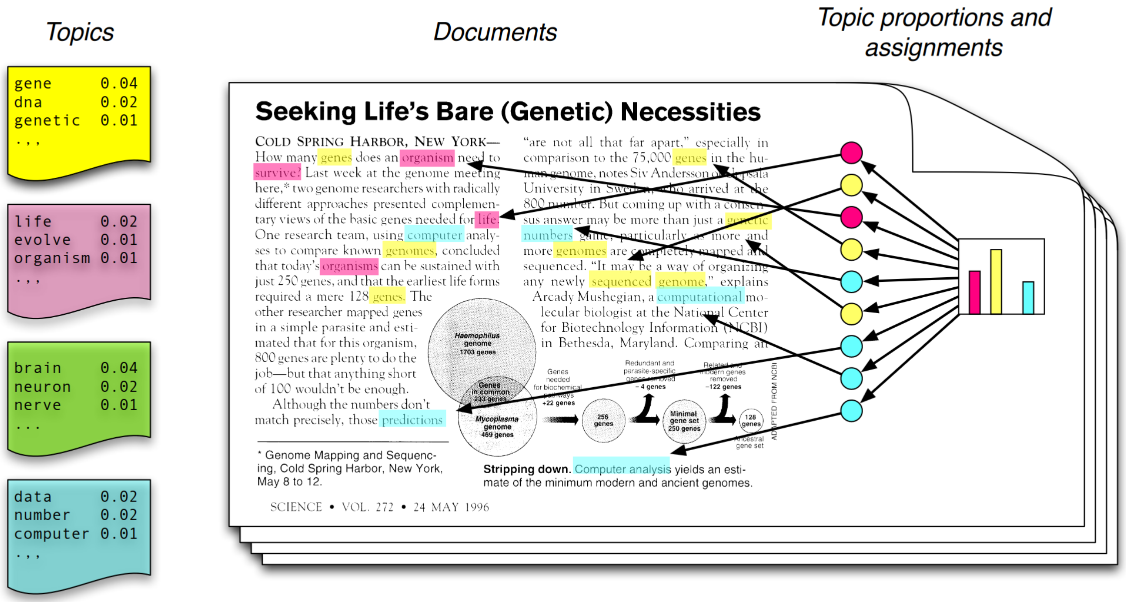
$$H_0 : S_{apartment} \geq S_{house}$$

```r
gumtree_sentiment_list <- gumtree_sentiment %>%
  filter(dwelling_type %in% c("apartment", "house"))

wilcox.test(sentiment ~ dwelling_type, data = gumtree_sentiment_list,
            alternative = "less")

#      Wilcoxon rank sum test with continuity correction
#
# data:  sentiment by dwelling_type
# W = 19358132, p-value = 0.0006765
# alternative hypothesis: true location shift is less than 0
```

Topic Modeling

Topic modeling is one of the core tools within Natural Language Processing (NLP). The goal of using topic modeling, is to assist the analyst in order to better segment large pieces of text into various clusters or "topics". A single piece of text will be a mixture of various topics with (hopefully) one of the topics being a dominant feature.

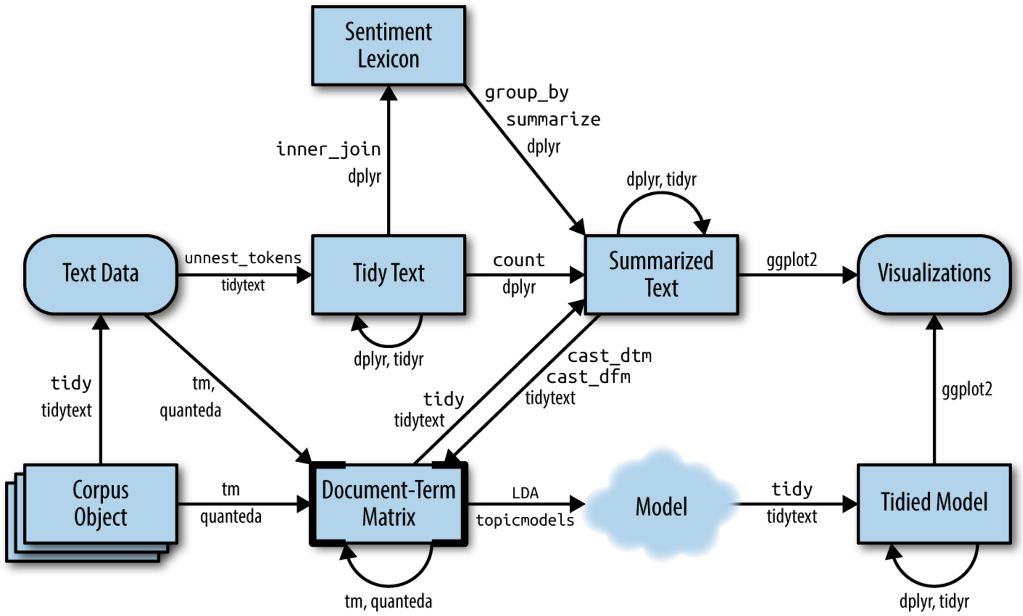
- The analyst has to make a subjective choice on the number of cluster
- Every document is a mixture of topics
- Every topic is a mixture of words



Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM, 55(4), pp.77-84.

Topic Modeling

A flowchart of a text analysis that incorporates topic modeling. The topicmodels package takes a Document-Term Matrix as input and produces a model that can be tided by tidytext, such that it can be manipulated and visualized with dplyr and ggplot2.



See *https://www.tidytextmining.com/topicmodeling*

Case Study: Farm descriptions

From tokens to DTM

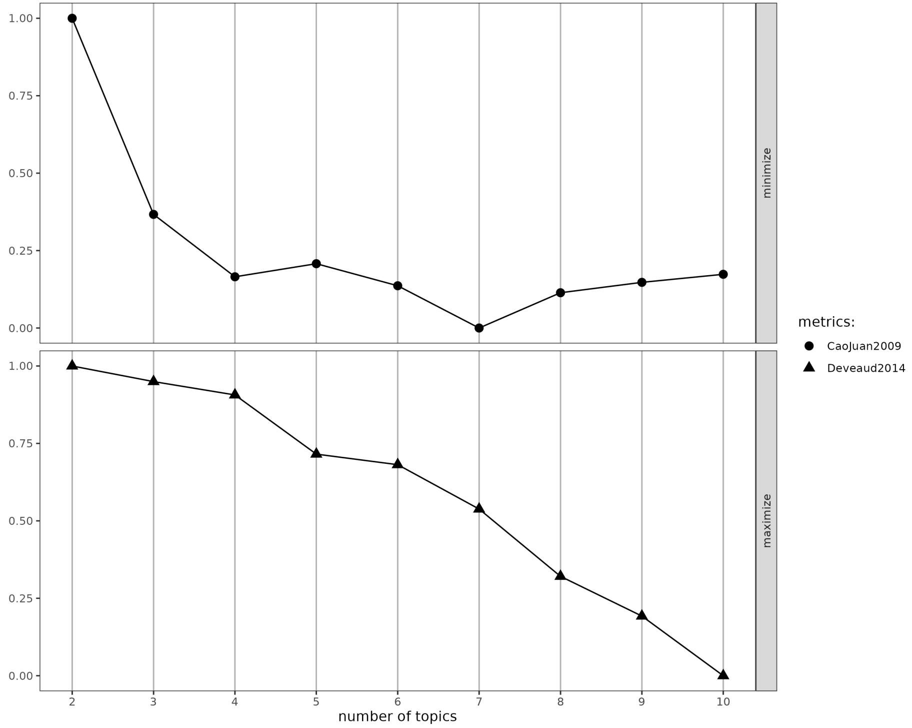
Topic models need to have some kind of design matrix: **DFM - Document Frequency Matrix** or **DTM - Document Term Matrix**. Luckily for us we already know how to get the count of terms per document!

```
gumtree_dtm <- gumtree_clean %>%
  filter(dwelling_type == "farm") %>%
  select(ad_id, description) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words, by = join_by(word)) %>%
  filter(!grepl("[0-9]+", word)) %>%
  count(ad_id, word) %>%
  cast_dtm(ad_id, word, n)

# <<DocumentTermMatrix (documents: 76, terms: 3306)>>
# Non-/sparse entries: 9306/241950
# Sparsity           : 96%
# Maximal term length: 34
# Weighting          : term frequency (tf)
```

How many topics?

In order to find out what K, or number of topics should be, is a bit more of an "art" than a pure science. One can start by looking at some statistics, but they are not absolutes and you will have to use your own judgement when conducting research:

```
library(ldatuning)
result <- FindTopicsNumber(
  gumtree_dtm,
  topics = seq(from = 2, to = 10, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)

FindTopicsNumber_plot(result)
```



Run the Model

We can then use the `LDA()` function to create a four-topic model. This is also mostly driven by theoretical hypothesis. My believe would be we should see: small holdings, citrus farms, wine farms and game farms... lets see if I am correct:

```
gumtree_lda <- LDA(gumtree_dtm, k = 4, control = list(seed = 1234))
gumtree_lda

# [main]>gumtree_lda
# A LDA_VEM topic model with 4 topics.
```

Now lets analyse the output:

- What words are within the topics?
- Prevalence of each topic in the corpus?

```
topics_beta <- tidy(gumtree_lda, matrix = "beta")
topics_gamma <- tidy(gumtree_lda, matrix = "gamma")
```

Understanding the topics

To understand the topics better we can analyse what words are most prevalent in a topic. This is called the `beta` matrix:

```
topics_beta
```

```
# [main]>topics_beta
# # A tibble: 13,224 × 3
#    topic term        beta
#    <int> <chr>      <dbl>
#  1     1 access 2.16e-  3
#  2     2 access 1.45e-  3
#  3     3 access 9.67e-  4
#  4     4 access 3.26e-  3
#  5     1 approx 5.78e-284
```

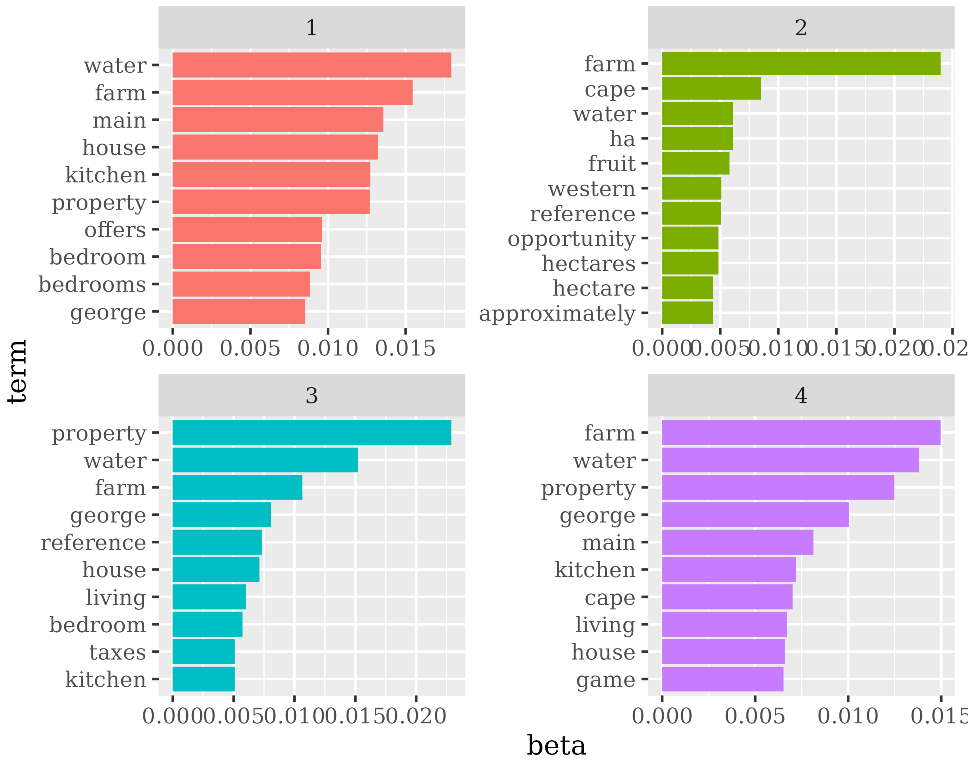
Lets use `slice_max` to find the topic 10 words per topic:

```
top_terms <- topics_beta %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

Understanding the topics

To understand the topics better we can analyse what words are most prevalent in a topic. This is called the `beta` matrix:

```r
library(ggplot2)

top_terms <- topics_beta %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



Topic Prevalence

To analyse how prevalent a given topic is in the corpus we use the `gamma` matrix or "topic probability per document". This tells us if a certain topic dominates or not:

```r
top_terms_group <- top_terms %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  summarise(top_words = paste0(term, collapse = ","))

topics_gamma %>%
  group_by(topic) %>%
  summarise(mean_gamma = mean(gamma)) %>%
  left_join(top_terms_group) %>%
  mutate(topic = glue("topic ({round(mean_gamma, 3)*100}%)")) %>%
  ggplot(., aes(reorder(topic, mean_gamma), mean_gamma,
                label = top_words)) +
  geom_col(fill = "#189bcc") +
  geom_label() +
  ylim(0, 0.45) +
  labs(x = "Topic", y = "Gamma") +
  coord_flip() +
  theme_minimal()
```