welcome

important

- this is a safe-to-fail experiment
- this is an exploration journey (and a pet project) for me
- all resources will be made available for experimentation/use
- any feedback, input and comments will be highly appreciated
- underpinning statistics will not be covered
- I will be using examples from development
- stay open-minded (there is a lot and some of it looks scary)!

agenda

- nature and purpose of sensemaker narratives
- [very] brief introduction to natural language processing
- examples: R Notebook, dashboard
- what to be aware of
- resources, what's next, discussion and q&a

terms

- sm sensemaker
- sf signification framework, sm data collection instrument
- self-signification the process of ‘coding’ the meaning by responding to a number of pre-defined questions
- nlp natural language processing
- GitHub a depository of codes open to all
- notebook allow you to interactively build narratives around small chunks of code and then publish the complete notebook as a report.

sm narratives (in a nutshell)

collection to minimise data bias

- by triggering memory of an event;
- to capture experiences that resonate most.

analysis

- to provide additional context;
- secondary to self-signification [!]

sm narratives what if?

…there are issues we did not think of: what else resonates with respondents?

…our assumptions about key concepts are incorrect: what terminology, which concepts people relate to?

…people talk about fundamentally different issues: depending on who they are

why nlp

- systematic analysis of textual data
- well researched and commonly used field
- works with non-english and multilingual texts
- we had access to UCL master students and faculty

nlp

- computer assisted text analysis
- based on machine learning - a method of data analysis, uses algorithms that iteratively learn from data, allows computers to find hidden insights without being explicitly programmed where to look
- about underlying semantic properties of text based on computational linguistics

step 1: dataset

project fragments of impact

organisaion undp

country Moldova

focus distribution of wealth, access to public services, inequalities caused by mass urbanisation

step 2: preparing data

create ‘corpus’

create ‘tokens’

remove digits and punctuation

removing stop words

notebook and dashboard

Code

NLP Webinar

Anna Hanchar, PhD

The Data Atelier
anna.h@thedataatelier.com

26 October 2017

notebook

Background reading

For a more general introduction to NLP in developing context see our recent paper "Data Innovation for International Development: An overview of natural language processing for qualitative data anlaysis" in proceedings of the Frontiers and Advances of Data Science IEEE Conference. Available from ArXiv https://arxiv.org/abs/1709.05563

There are a couple of good, introductory overview papers about NLP in political science literature. For example, Lucas et al. "Computer-Assisted Text Analysis for Comparative Politics", Political Analysis, 2015, 23: 254-277. Available here: http://christopherlucas.org/files/PDFs/text_comp_politics.pdf

Grimmer and Stewart provide a good overview in their "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", Political Analysis, 2013. Available here: http://web.stanford.edu/~jgrimmer/tad2.pdf

We are extensively using the "quanteda" package. For an introductory tutorial see: https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

We use R Markdown and R Notebooks for completely transparent and reproducible research. For an introduction to R Markdown and R Notebooks environment please see: http://rmarkdown.rstudio.com and http://rmarkdown.rstudio.com/r_notebooks.html

Loading packages and data

Load packages (if necessary need to be installed beforehand).

Hide

```
library(readtext)
library(quanteda)
library(dplyr)
library(stringr)
library(ggplot2)
library(haven)
library(readxl)
library(magrittr)
library(stm)
library(readr)
```

Loading data from original CSV file, specifying which column contains stories. Here, we're using a convenience function "readtext" from the eponymous package that simplifies the loading of text data. For more information see https://github.com/kbenoit/readtext.

Hide

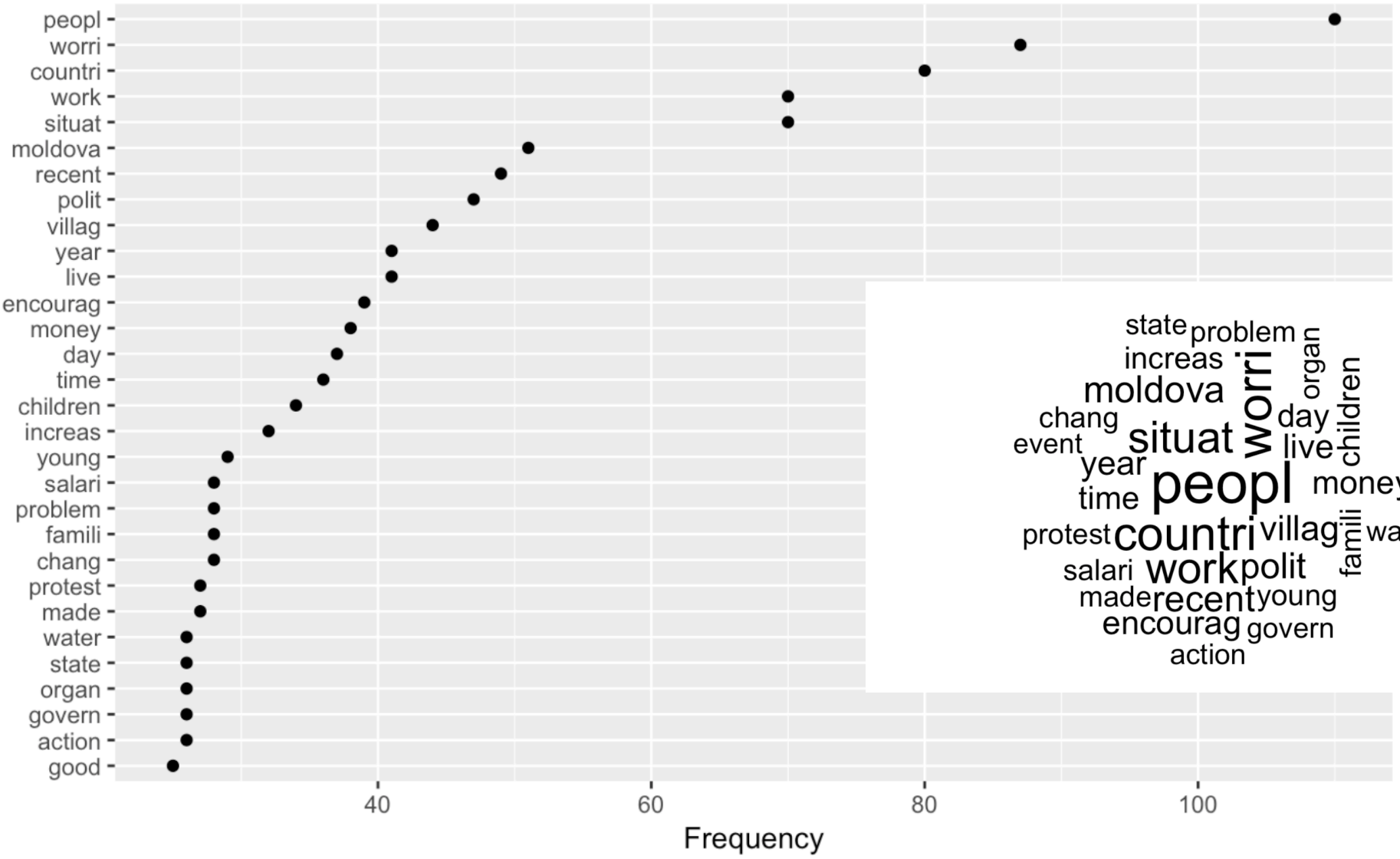
```
moldova_foi <- readtext("foimoldova2015_Standard.csv",text_field = "Your experienc
e")
```

Pre-processing data

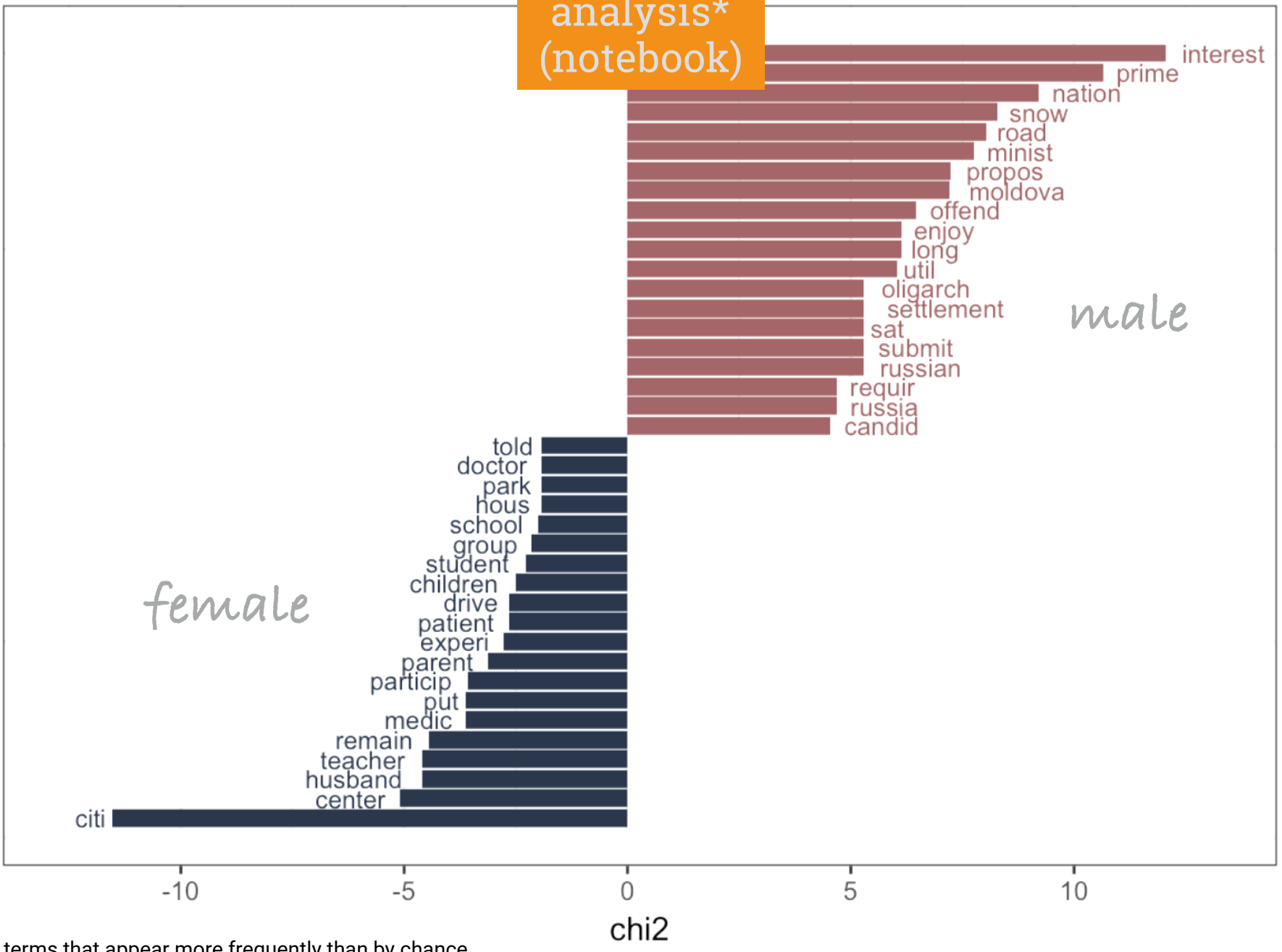
Once the data is loaded the next step is to create a "corpus" object for analysis

dashboard

step 3.1: frequency analysis (notebook)



step 3.2: keyness analysis* (notebook)



* shows key terms that appear more frequently than by chance

keyness: the frequency of a word in the text when compared with its frequency in a reference corpus

step 3.3: word and document similarity (notebook)

```
work_simil <- textstat_simil(dfm.trim, "work", method = "cosine", margin = "features")
as.list(work_simil, n = 15)
```

```
$work
   employ    depart    reduct    retire   appreci       job    worker    social
0.3600411 0.3402069 0.2886751 0.2777778 0.2721655 0.2608360 0.2566001 0.2364027
  compani     field      hire   minimum     retir     engag     month
0.2236068 0.2222222 0.2222222 0.2222222 0.2182179 0.2151657 0.2100420
```

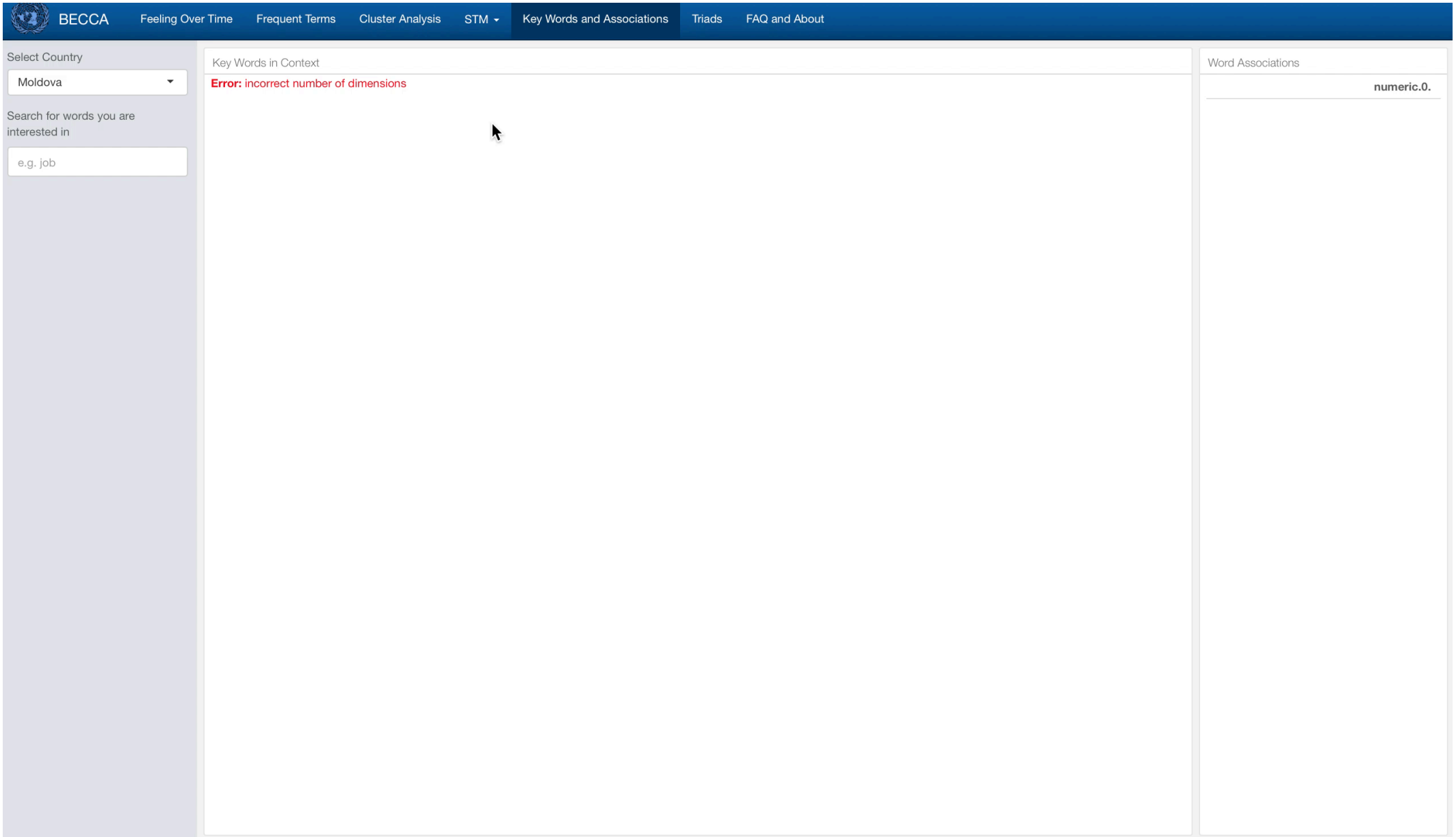
```
work_dist <- textstat_dist(dfm.trim, "work", method = "euclidean", margin = "features")
as.list(work_dist, n = 15)
```

```
$work
   peopl  countri   villag    worri   situat    water  moldova    polit encourag
15.16575 14.96663 13.71131 13.45362 13.34166 13.19091 13.07670 13.07670 12.44990
   money     road   recent    mayor   offend  student
12.32883 12.28821 12.12436 12.04159 12.04159 12.00000
```

Step 3.4: keyword in context (notebook)

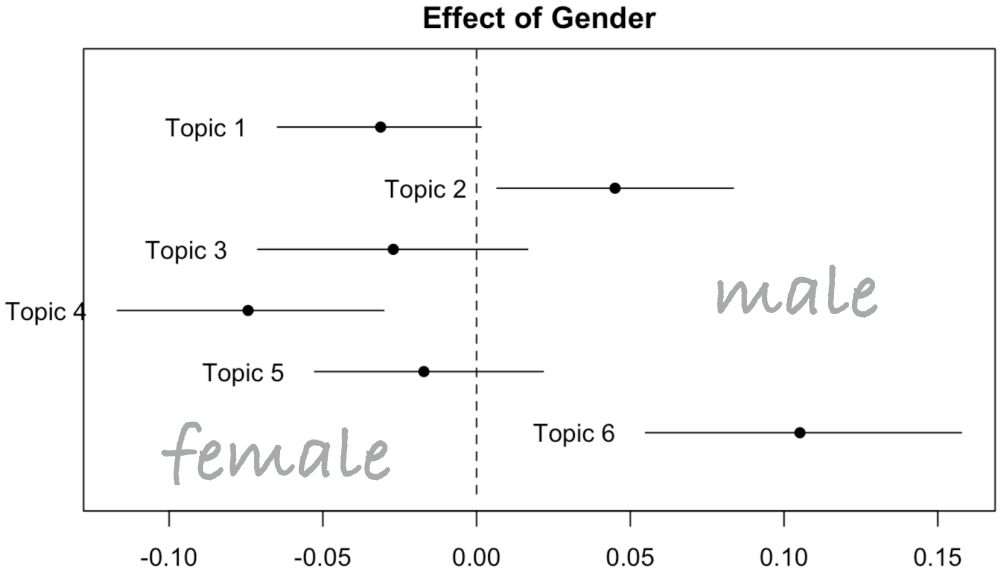
```
 [text1, 29]     aroused the interest of the |  work   | and the future development of
 [text2, 76]          turned out he wants to |  work   | but too many people are
 [text18, 3]                        I receive | salary  | in lei. And the
[text18, 35]   in relation to currencies and | salary  | remains the same. If
 [text26, 6]      nehochu go abroad but this |  work   | is not very worried
[text40, 20]                  . Just a lot of |  work   | yard by the end of
 [text47, 2]                        Changing |  jobs   | raised concerns me if it
[text57, 29] retirees others were allowed to |  work   | release; workers with high
[text57, 32]        allowed to work release; | workers | with high stint working in
[text57, 36]       ; workers with high stint | working | in a field were moved
```

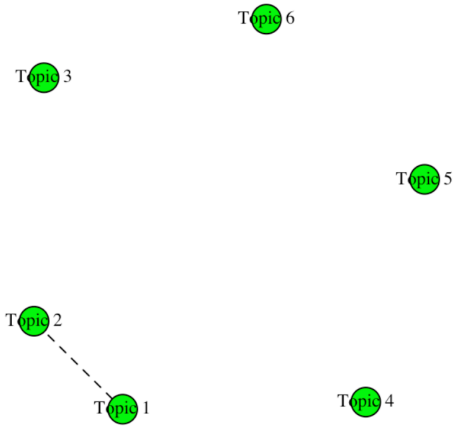
Step 3.4: keyword in context (dashboard)



Step 3.5: topic modelling (notebook)

```
Topic 1 Top Words:
     Highest Prob: peopl, worri, increas, live, children, price, pay
     FREX: peopl, price, affect, care, util, product, mother
     Lift: carpet, casino, requir, sum, affect, price, accid
     Score: peopl, price, increas, femal, small, employe, affect
Topic 2 Top Words:
     Highest Prob: work, money, young, pension, social, man, abroad
     FREX: money, pension, social, man, abroad, law, activ
     Lift: activ, law, pension, russia, social, visit, man
     Score: work, money, pension, social, young, paid, man
Topic 3 Top Words:
     Highest Prob: villag, road, year, time, school, job, left
     FREX: road, job, decid, experi, found, husband, learn
     Lift: abandon, assist, assum, balti, basi, block, brutal
     Score: road, job, employ, villag, colleg, water, wait
Topic 4 Top Words:
     Highest Prob: encourag, recent, citi, made, particip, mayor, feel
     FREX: particip, local, build, fact, park, rule, apart
     Lift: alley, build, construct, creativ, discuss, donat, inherit
     Score: snow, project, encourag, citi, build, essay, republ
Topic 5 Top Words:
     Highest Prob: protest, event, day, remain, organ, year, hous
     FREX: protest, event, remain, result, center, indiffer, room
     Lift: adult, alexand, approach, area, assembl, bag, basement
     Score: sister, offend, dog, protest, aggress, event, water
Topic 6 Top Words:
     Highest Prob: countri, moldova, situat, polit, worri, chang, govern
     FREX: countri, moldova, polit, govern, minist, elect, moldovan
     Lift: countri, govern, accept, affair, airport, alegiril, arous
     Score: polit, countri, moldova, minist, prime, govern, billion
```





Step 3.5: topic modelling (dashboard)

BECCA | Feeling Over Time | Frequent Terms | Cluster Analysis | STM | Key Words and Associations | Triads | FAQ and About

Select Country

Moldova

Select Topic

Topic 1

Topic Labels

```
Topic 1 Top Words:
         Highest Prob: encourag, action, peopl, made, communiti, think, chisinau, feel, citizen, organ, recent, day, center, event, hospit, person
         FREX: encourag, chisinau, communiti, action, made, center, environ, necessari, hospit, differ, citizen, organ, feel, idea, volunt, think
         Lift: almost, centr, environ, music, necessari, patriot, solidar, status, style, youth, chisinau, block, volunt, encourag, syslab, communiti
         Score: abandon, absenc, accept, accord, admir, age, aid, alon, alreadi, anim, announc, appeal, appoint, arriv, aspect, assembl
```

Topic Wordcloud

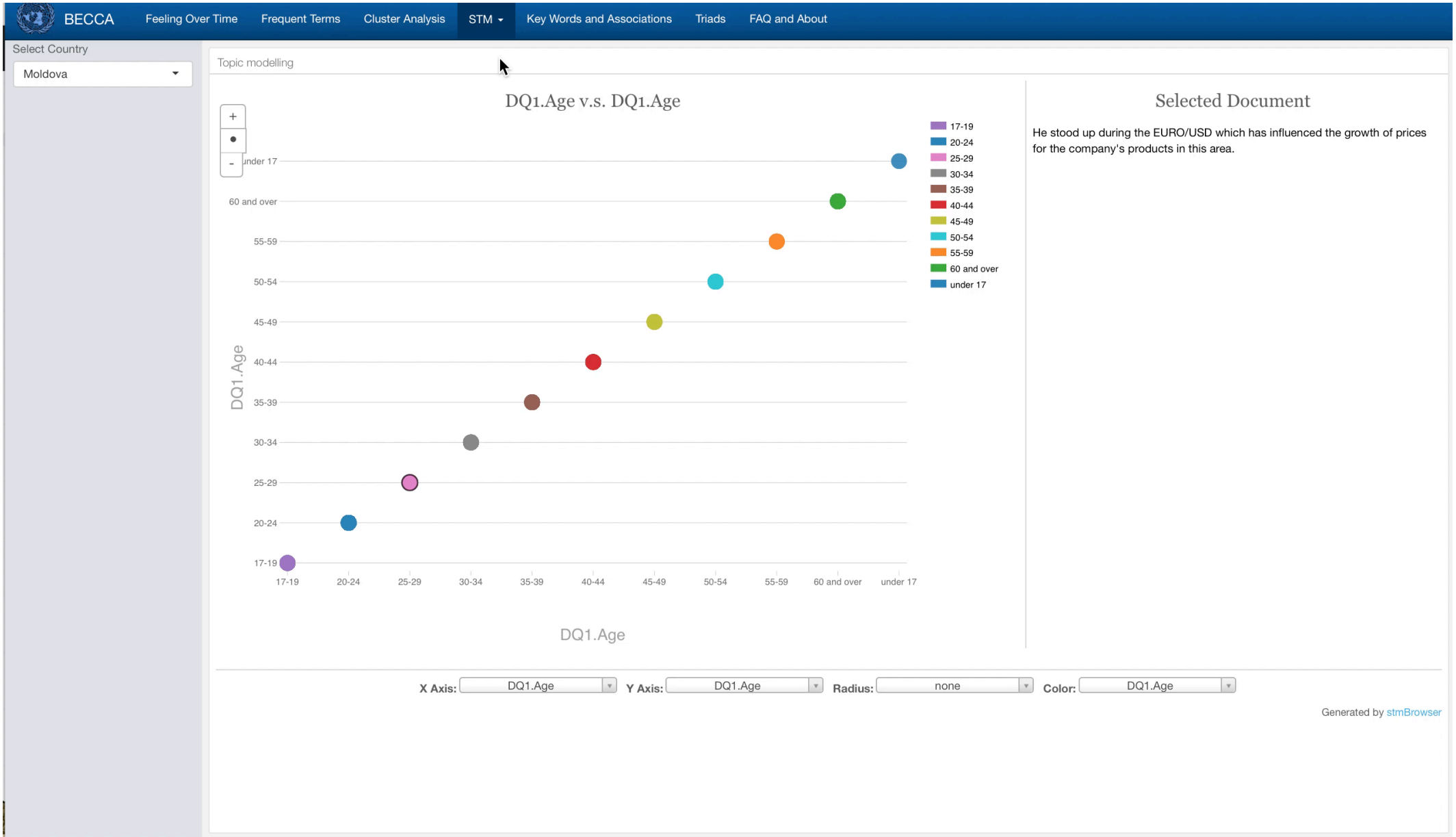
Most Representative Stories

Recently I've learned of the Syslab centre and I signed up. Here every day people we greeted with a smile and with sincerity in what ïnterprind. At the center I feel encouraged by these people because they make me feel much more optimistic and much stronger than I am. I met you a day in the ask I felt offended by someone about the specialization that I mentee and these people especially at syslab encouraged me to fight for what they are because each of us have strengths different.

In the summer of 2015 in the center of Tiraspol, I like many others have heard every day to the classical music coming from the speakers. But the essence of the other. There was a comical situation. Center. Quay. People in sports clothes click the seeds on the pavement drinking a liter bottle of beer behave uncivilized. And amid all this picture plays Bach.The picture is small.

I'm in Transnistria due to the contribution of the EU have rebuilt 2 hospitals. I'm encouraged by the idea that we may be able to join the RM with MRI.

Step 3.5: topic modelling (dashboard)



what to be aware of...

- sf design: If NLP to be used, quality of narratives is important, resources should be available for transcription (if audio is collected)
- study design: preparation to and application of nlp
- analysis: self-signification takes priority and NLP should only be used as an additional approach to make sense of data

resources

- slides
- notebook (GitHub): papers, explanations, codes, outputs
- myself (anna.h@thedataatelier.com)

what's next

- notebook: ready to use; can be modified, we ask for all modifications to be shared on github
- dashboard: needs work; we will share once done

q&a

- thoughts, ideas, feedback, suggestions...