Hackathon on data science for STI policy

STIP Lab and OECD – TIP event

Research question:

To what extent is it possible to characterise typologies of policy proposals on the theme of scientific employment and research careers?

25 June 2022

SPRU team

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

EQUIS ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Objectives and Agenda

Objectives

To identify key themes related to scientific employment and research careers in available policy proposal databases, so to support the decision-making process

Methodology:

1. Using the label of ‘theme’ to narrow down research objective (TH44_Inter-sectoral mobility, TH53_Research careers, TH54_Gender balance and inclusiveness)
2. Applying multiple basic techniques to conduct clustering analysis and comparing the results
3. Combining literature review and qualitative analysis

Agenda

1. Data pre-processing
2. Descriptive analysis
3. Identifying the cluster
4. Discussion

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Data pre-processing

Key points

1. Following the instruction of ***Getting Started with NLP of Research and Innovation Policy Data using R*** given by OECD
    1. *Preparation: load R packages and download data*
    2. *Prepare the dataset*
    3. *Prepare and pre-process textual data*

2. Get the textual data from STIP dataset
   'Description' + 'Objectives' columns

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Descriptive analysis

Key points

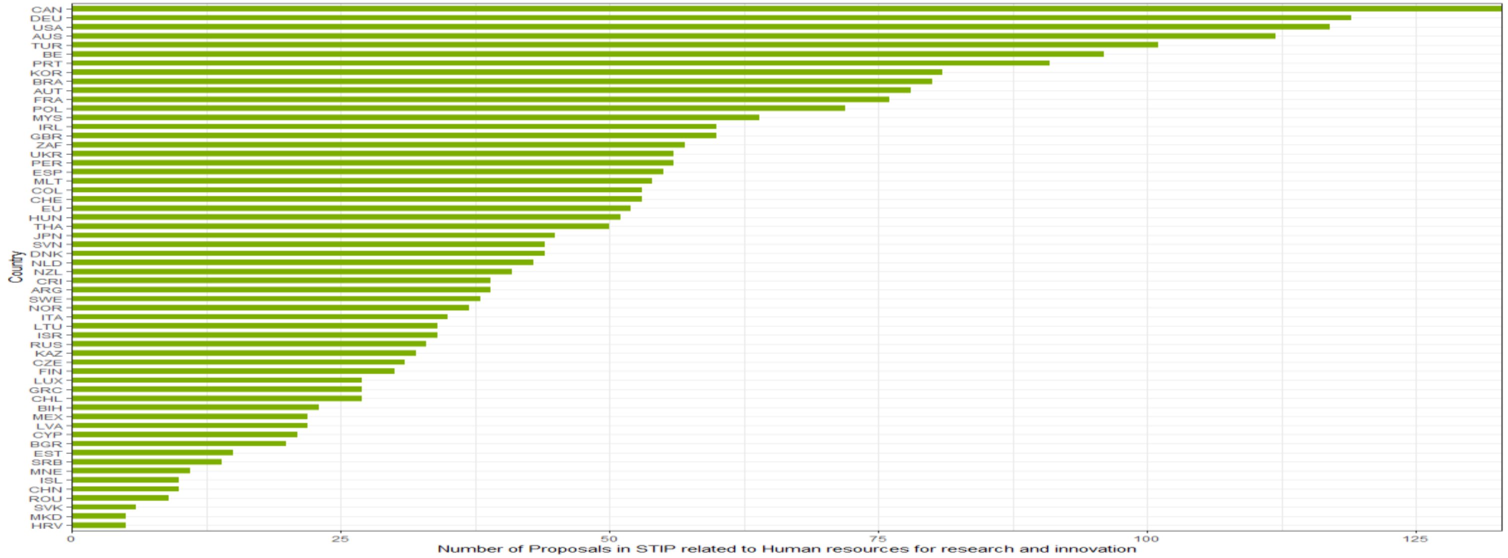
1. Showing the number and percentage of policy proposal related to HR policies by bar chart and heat map

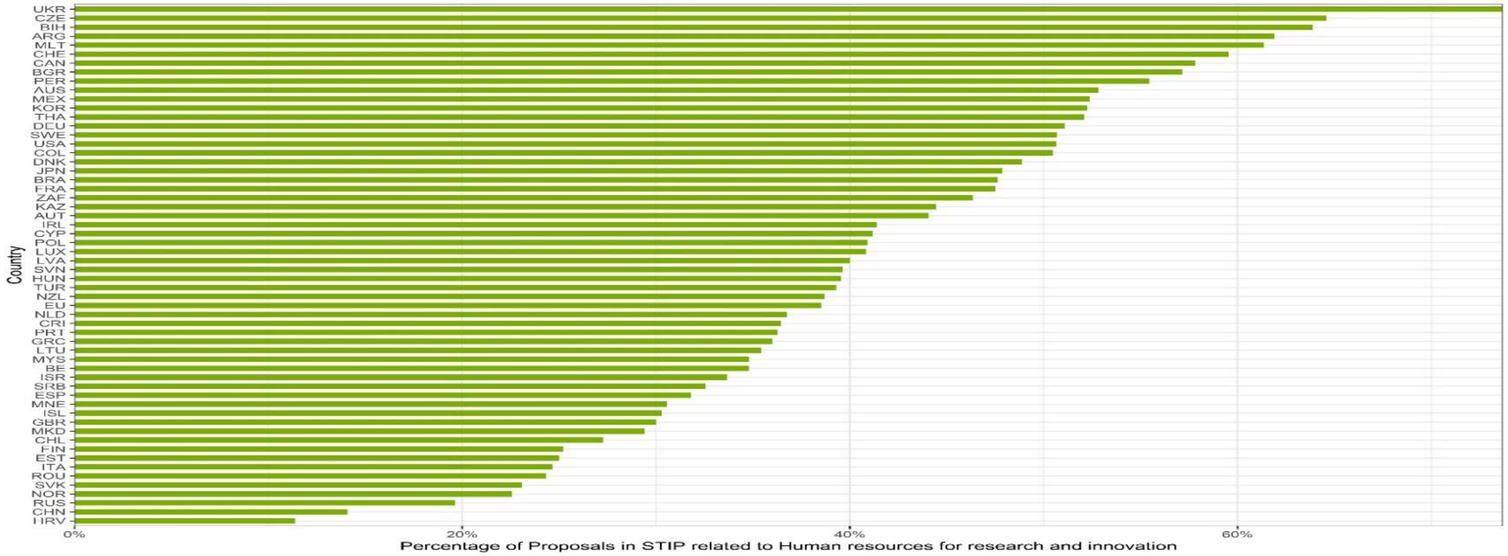
2. Showing the budget-weighted heat map

3. Showing proposal distribution across themes

Proposals related to HR for R&I per country

57 countries

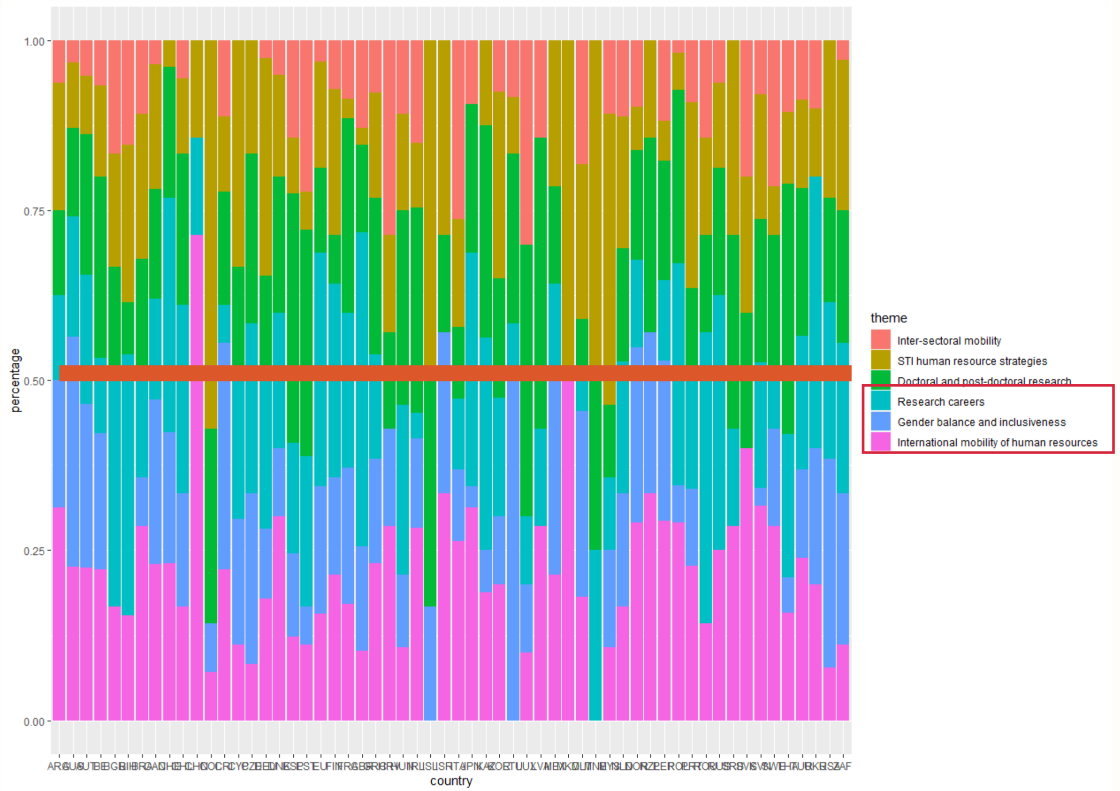




Key questions:

1. The number of proposals are quite various across countries
2. Analysis sample is unbalanced when we compare the textual data across countries

Budget Heatmap and distribution



Key questions:

- Difference in comparison between the distribution of budget-weighted and number-weighted
  - Budget: Doctoral and post-doctoral research, International mobility of human resources, STI human resource strategies
  - Number: International mobility, Research careers, Gender balance and inclusiveness

Identifying the cluster

Clustering analysis (PCA, Hierarchical Clustering, K-mean Clustering)

1. Extract the text of HR policies

   (Extract text information is from “ShortDescription” and all “Objectives” in the dataset)

2. Delete words according to frequency (the most and the least used), delete meaningless words manually

   (e.g. one, two, also…)

3. Get tfidf (get a term frequency of country-term matric like)

4. Conduct PCA analysis (2 components that explain around 20% information of the whole text.)

5. K-means Clustering

   1. Select k value: Elbow method, Average silhouette method, Gap statistic method, PCA approach and hierarchal clustering
   2. Hierarchical Clustering
   3. Conduct k-means clustering, compare results by term frequency in each cluster
   4. Connect results with literature review

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Identifying the cluster

Key questions:

1. Aggregating textual data by using country as an analysis unit (unbalanced dataset)

Terms

```
Document-feature matrix of: 50 documents, 341 features (74.78% sparse) and 27 docvars.
      features
docs   fund opportun intern team erc grant support plan appli academ
  ARG     0        0      0    0   0     0       0    0     0      0
  AUS     1        3      1    1   0     1      12    2     1      0
  AUT    16        1      7    3   4     2      13    6     4      1
  BE      2        0      0    0   0     0       2    0     1      2
  BGR     5        0      0    0   0     0       0    0     0      1
  BIH     1        0      2    0   0     0       3    1     0      0
[ reached max_ndoc ... 44 more documents, reached max_nfeat ... 331 more features ]
>
```

Countries

Identifying the cluster

Key questions:

2. Principle component analysis:

How to increase the information included in two components?

- Topic modeling to reduce the dimensions of the data?
- Manually delete the dimensions of the data (useless words)?

Th 53 Research Career

PCA based (23%)

R package: fviz_cluster

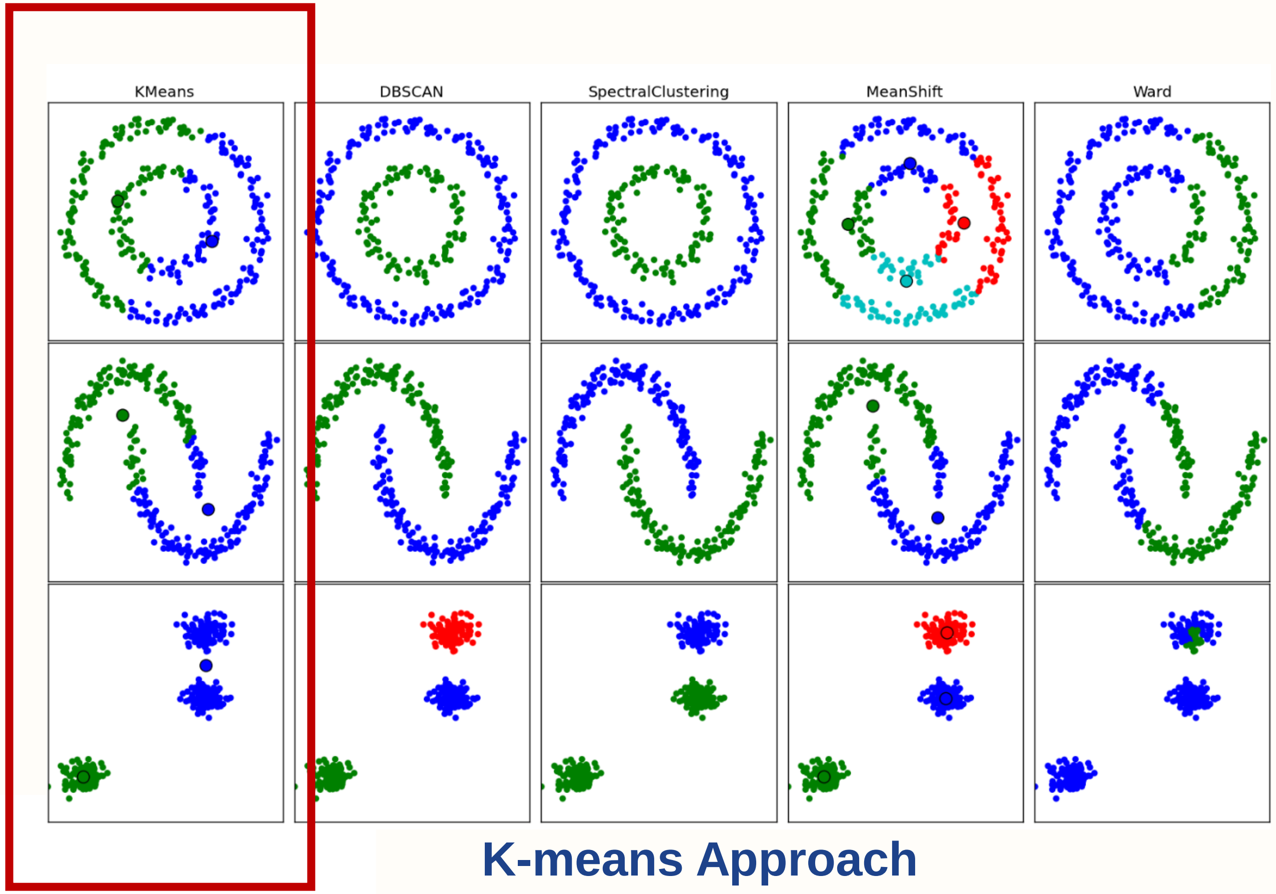
Identifying the cluster

Key questions:

3. K-means vs Hierarchical Clustering:

How to identify textual data structure that are suitable to apply these approaches?

- K Means clustering is found to work well when the structure of the clusters (like circle in 2D, sphere in 3D) is hyper spherical



K-means Approach

Identifying the cluster

Key questions:

4. Bigram vs unigram:

From terms:
Bigram approach makes more sense

From k-mean and PCA:
Unigram approach makes more sense

```
> dfm_countries@Dimnames$features
 [1] "reform"       "program"      "creation"     "nation"       "system"       "univers"
 [7] "scientif"     "develop"      "field"        "role"         "contribut"    "strengthen"
[13] "improv"       "technolog"    "product"      "establish"    "criteria"     "evalu"
[19] "activ"        "countri"      "qualiti"      "teach"        "staff"        "state"
[25] "recruit"      "phd"          "graduat"      "line"         "prioriti"     "ministri"
[31] "educ"         "scienc"       "innov"        "programm"     "futur"        "women"
[37] "stem"         "leader"       "scholarship"  "partnership"  "industri"     "support"
[43] "skill"        "particip"     "job"          "scientist"    "impact"       "engin"
[49] "address"      "respons"      "review"       "train"        "carri"        "ensur"
[55] "meet"         "need"         "higher"       "degre"        "divers"       "strategi"
[61] "enabl"        "differ"       "peopl"        "potenti"      "world"        "build"
[67] "cultur"       "work"         "action"       "plan"         "earli"        "advanc"
[73] "set"          "foundat"      "approach"     "achiev"       "sustain"      "increas"
[79] "gender"       "chang"        "govern"       "practic"      "lead"         "career"
```

unigram

```
> dfm_countries@Dimnames$features # look at the words and look for the words that you want to delete
 [1] "nation_system"        "univers_research"    "scientif_research"   "technolog_activ"
 [5] "research_system"      "na_na"               "teach_staff"         "research_train"
 [9] "nation_scienc"        "action_plan"         "gender_equiti"       "earli_career"
[13] "career_research"      "support_research"    "research_project"    "appli_research"
[17] "intern_research"      "research_collabor"   "prioriti_area"       "higher_educ"
[21] "educ_institut"        "career_develop"      "career_stage"        "provid_support"
[25] "research_institut"    "research_sector"     "encourag_research"   "programm_support"
[29] "provid_financi"       "financi_support"     "equal_opportun"      "excel_research"
[33] "outstand_research"    "young_research"      "erc_grant"           "research_career"
[37] "research_area"        "public_sector"       "work_condit"         "promis_research"
[41] "research_team"        "feder_govern"        "austrian_scienc"     "scienc_fund"
[45] "three_year"           "research_fund"       "best_research"       "basic_research"
```

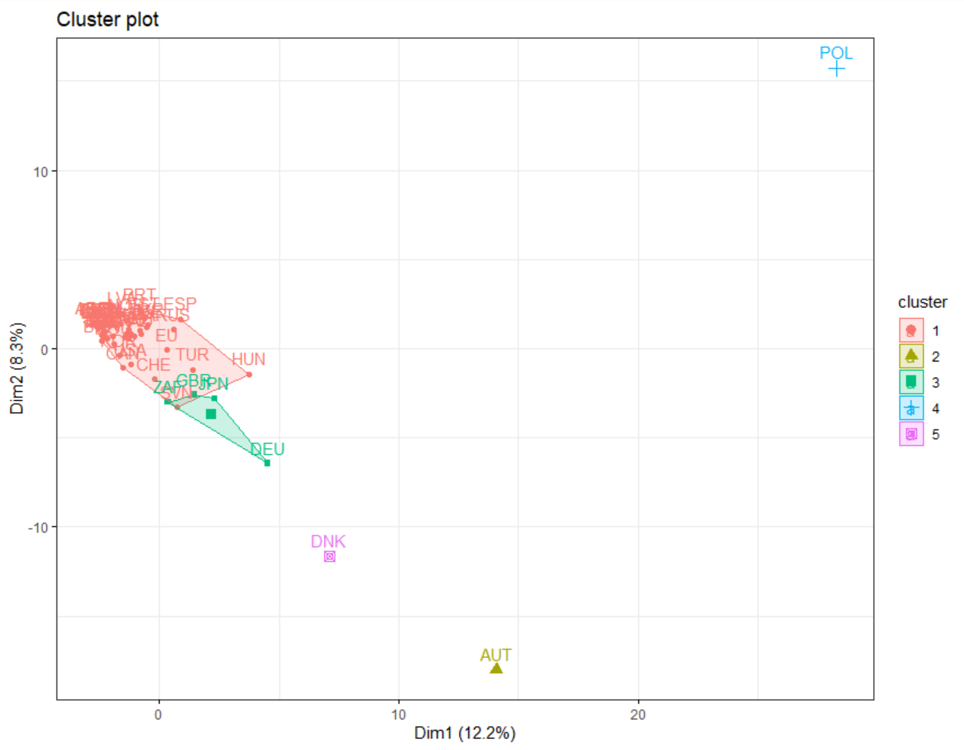
Bigram

Identifying the cluster

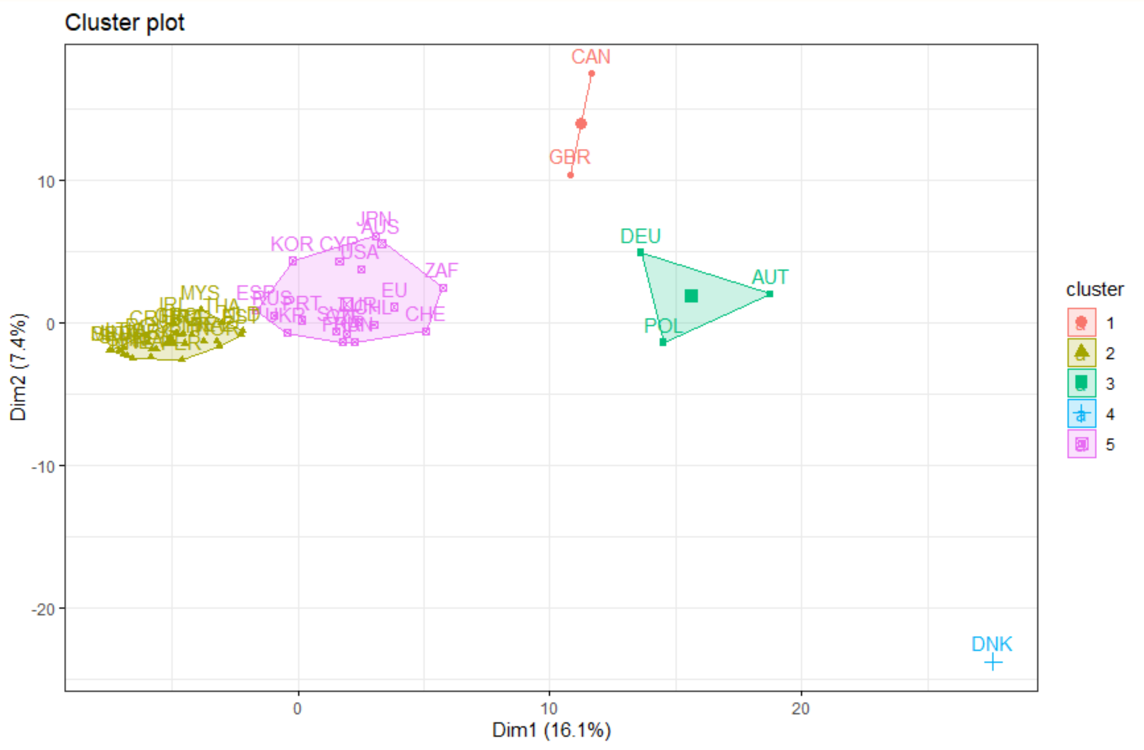
Key questions:

4. Bigram vs unigram:

Th 53 Research Career



Bigram



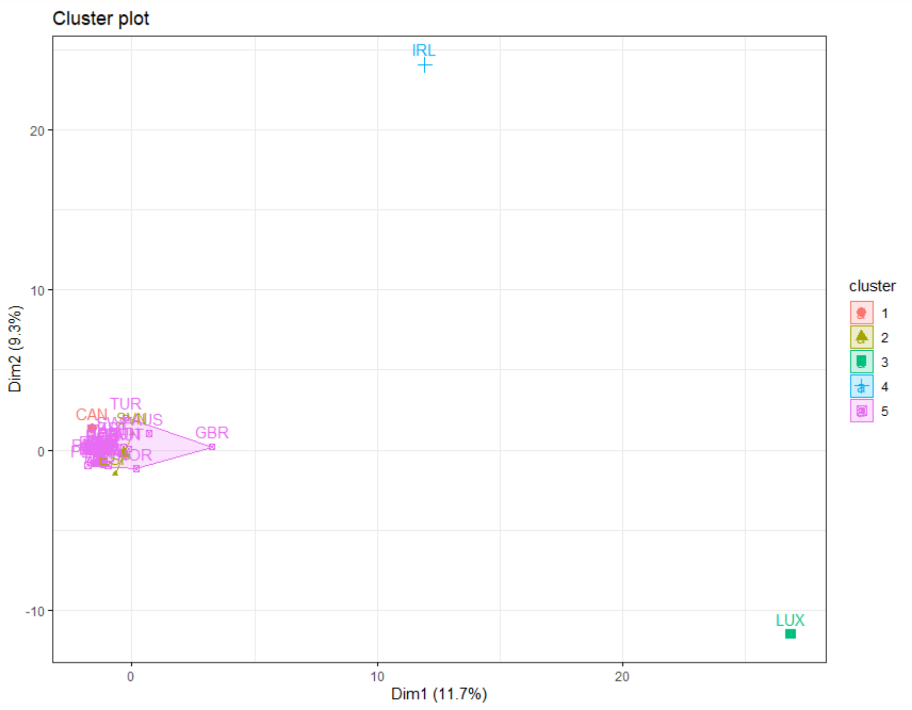
unigram

Identifying the cluster

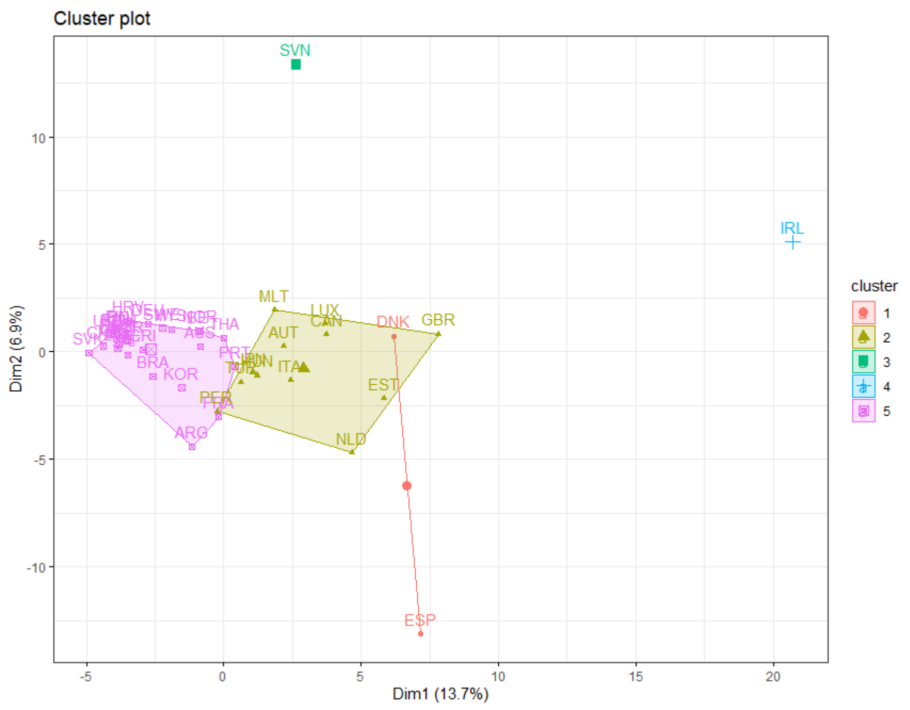
Key questions:

4. Bigram vs unigram:

Th 53 Research Career



Bigram



unigram

Conclusion

- The number of proposals are quite various across countries. Analysis sample is unbalanced when we compare the textual data across countries.
- Difference in comparison between the distribution of budget-weighted and number-weighted, and how to use budget-weighted information to conduct clustering analysis?
- How to increase the information included in two components?
- How do you identify textual data structure that are suitable to apply these approaches (K-means vs Hierarchical Clustering)?
- How to conduct clustering analysis combining the bigram approach? How to understand the role of bigram approach in clustering analysis?

Thank you – The SPRU team

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL