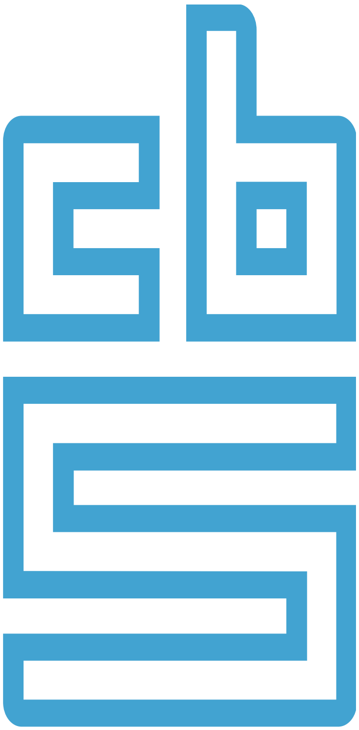


Adjusting text-analysis to source motive

Paul Keuren, Statistics Netherlands

Sources

- Internet
  - Webpages
  - Social media
- Questionnaire (form)
  - Required input
  - Optional input
- Questionnaire (in person/telephone)
  - Required input
  - Optional input
  - Adjustments by interrogator

Motive

Where does the text come from?

- Person
- Generated

What was the goal of the text?

- Inform
- Activate
- …

Connecting Motives and Sources

Company pages

-> Advertising

Optional Question in questionnaire

-> Informative

-> Express annoyance

-> Meta question (questionnaire related)

Social Media

-> Advertising

-> Inform

-> Trolling

A Case of NACE

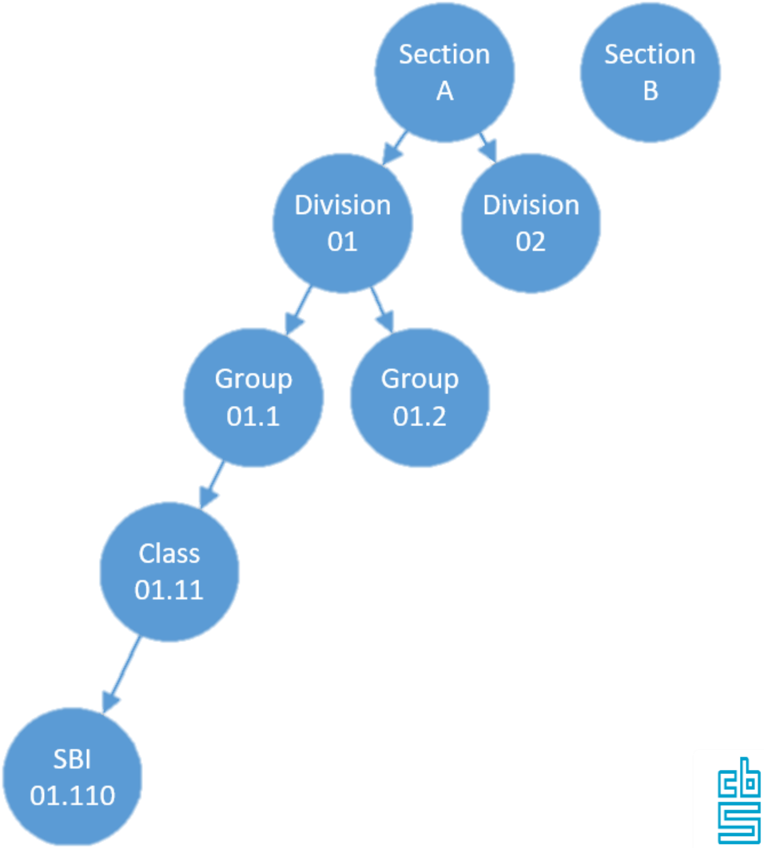
NACE & SBI

What is NACE?

Levels of NACE:

- 21 Sections (A-U)
- 88 Divisions (01-99)
- 272 Groups (0 – 9)
- 615 Classes (0 – 9)

SBI adds 1 level



Data Source

Chamber of Commerce (CoC)

CoC Employee asks questions Company answers

- Question of SBI is required
- Employee productivity = # registrations per time unit
- Certain SBI codes result in additional taxes

Some expected effects

- Copy paste of company descriptions from default SBI list
- Companies get advised on how to describe themselves

What could we want from it?

Automatic selection of SBI based on text

Extend current Semantic web Search tool

Precursor for fully automated SBI detection from webpages.

Analysis

1. Filter default texts (from SBI list)
2. Tokenize texts
3. Calculate binding between SBI and tokens
4. Evaluate tokens for Semantic web addition

5. Analyse SBI code system using the binding

Binding between token and code

Normalized (pointwise) mutual information*

If a token occurs for a score of:

-1, the code does not occur

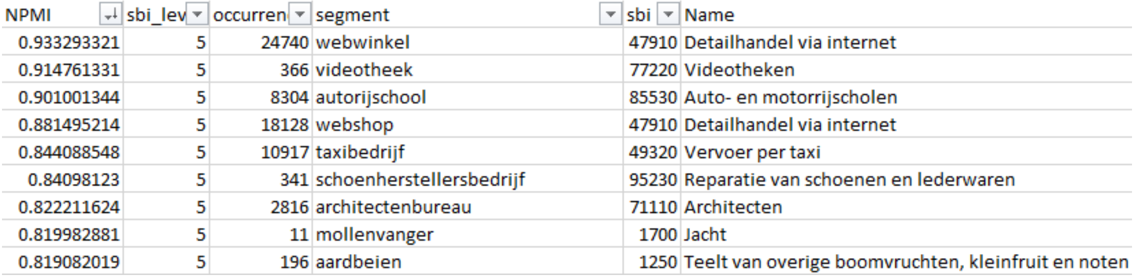
0, the code has a random chance to occur

1, the code occurs

Construct a sparse co-occurrence matrix (only code co-occurrences that occur)

*BOUMA, Gerlof. Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL, 2009, 31-40.

Results



NPMI	sbi_lev	occurren	segment	sbi	Name
0.933293321	5	24740	webwinkel	47910	Detailhandel via internet
0.914761331	5	366	videotheek	77220	Videotheken
0.901001344	5	8304	autorijschool	85530	Auto- en motorrijscholen
0.881495214	5	18128	webshop	47910	Detailhandel via internet
0.844088548	5	10917	taxibedrijf	49320	Vervoer per taxi
0.84098123	5	341	schoenherstellersbedrijf	95230	Reparatie van schoenen en lederwaren
0.822211624	5	2816	architectenbureau	71110	Architecten
0.819982881	5	11	mollenvanger	1700	Jacht
0.819082019	5	196	aardbeien	1250	Teelt van overige boomvruchten, kleinfruit en noten

Results usable for semantic web

Weights might also be usable for semantic web approach

Analyse SBI code system

Assume:

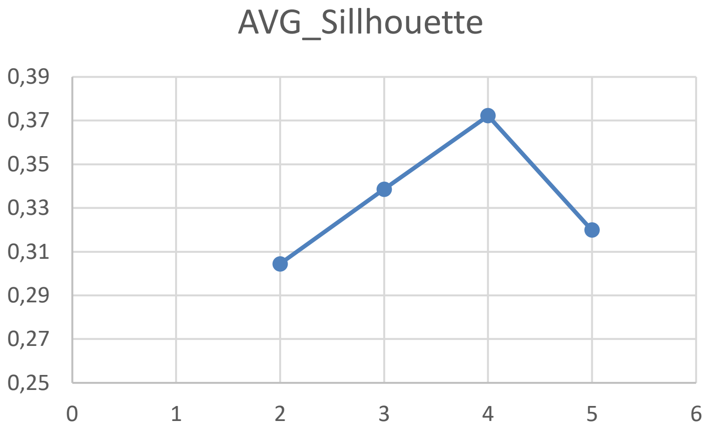
SBI codes = clusters

1 – Binding = distance to cluster

For each level:

- Use assumptions
- Compute the Silhouette score
- Use average Silhouette score as a quality

Evaluation



- Increase of binding with each NACE
- Decrease on additional SBI level

Facts that matter