

# Content analysis and natural language processing

CS-4750 'Introduction to Natural  
Language Processing', October 12, 3-  
3:50PM, EN-1001

# Key points

- Objectivist and subjectivists approaches
- Content analysis: a third way?
- Content analysis and its types (qualitative, quantitative & mixed methods)
- Assessing the validity and reliability of the outcomes: mixed methods content analysis
- *ThinkMate* online platform and possible topics for course projects

## Objectivist approach

- Words are viewed as having ‘proper senses’
- One can understand an object entirely in terms of a *set* of its *inherent* properties; the ‘truth’ is objective
- For objectivists the word *digest* would have two different and distinct literal (objective) meanings – *digest*<sub>1</sub> for food and *digest*<sub>2</sub> for ideas
- The analysis of other words with which the word *digest* co-occurs is indicative of whether it has to be interpreted as *digest*<sub>1</sub> or *digest*<sub>2</sub>

## Subjectivist approach

- Words are viewed as having senses attributed to them by the reader or listener
- Understanding of an object is always subjective; any ‘truth’ is relative and subjective
- For subjectivists, the word *digest* would have as many meanings as there are readers or listeners, *digest*<sub>*i*</sub>
- It is impossible to discriminate between various meanings

# Searching for a 'third way': Experientialism

- Lakoff, George & Johnson, Mark (1980), *Metaphors We Live By*. Chicago and London: The University of Chicago Press
- We understand experience metaphorically when we use concepts from one domain of experience (most often, our interaction with the physical environment) to structure experience in another domain. E.g., The concept 'in' emerges from spatial experience. The same goes for 'at' etc.
- From the experientalist point of view, being objective is always relative to a conceptual system

# Searching for a 'third way': Content analysis

- **Content analysis:** the analysis of texts or images (but today I'll focus on texts only) as a source of information about social action. It consists in 'mining' the text/image for relevant data. 'Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use' (Krippendorff, Klaus, 2004, *Content Analysis*, Sage, p.18). My own definition of content analysis: a study of how readers interpret and comprehend texts/images
- Content analysis is a relatively 'neutral' research method that can be adapted to the needs of discourse analysis, critical discourse analysis, visual research methods and some other theories and research methods, including natural language processing (especially **quantitative** content analysis)

NESS has to do with form, while the STRENGTH OF EFFECT has to do with meaning. Thus the metaphor CLOSENESS IS STRENGTH OF EFFECT, which is part of our normal conceptual system, can work either in purely semantic terms, as in the sentence "Who are the men closest to Khomeini?," or it can link *form* to *meaning*, since CLOSENESS can indicate a relation holding between two *forms* in a sentence. The subtle shades of meaning that we see in the examples given above are thus the consequences not of special rules of English but of a metaphor that is in our conceptual system applying naturally to the *form* of the language.

→ overstated & confused\*  
 \* Oh, give it up, buddy, this is a Library book.

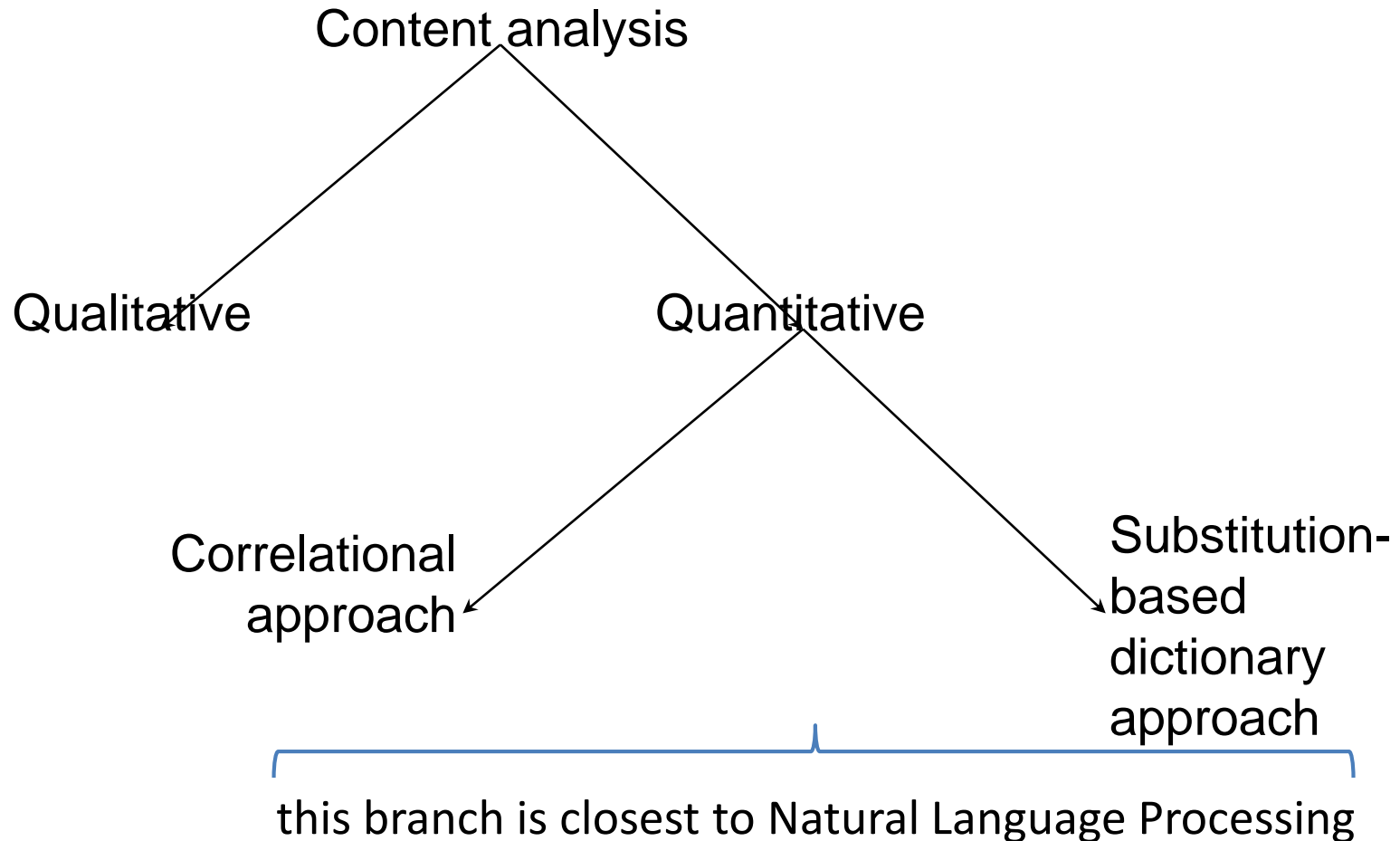
The ME-FIRST Orientation

Save the meta commentary for your own frigging books.  
 J. 002.

Comments on the QEI copy of Lakoff & Johnson's book (at p. 132):

- The sentence 'the subtle shades...' is deemed to be 'overstated & confused' by one reader
- The other reader rebukes: 'oh, give it up, buddy, this is a library book. Save the meta commentary for your own frigging books'

# Types of content-analysis



# Qualitative content analysis

- **In vivo** coding: identifying relevant (to a particular research question) and simply interesting fragments of a text, image 'from scratch', i.e. the list of codes is an outcome of the coding
- Coding as manual attribution of particular meanings to fragments of a text with the help of a **codebook**. This type of qualitative content analysis requires
  - Clearly defined concepts (**Codes**) and rules for their application
  - Reliance on the coder's subjective judgment as to how to apply these rules (to reduce its subjectivity, coding is sometimes done in teams → the coefficient of inter-coder agreement, such as Krippendorff's alpha)



# Quantitative content analysis

- **Correlational analysis:** the analysis of co-occurrences of words. It helps identify words that tend to co-occur in the text and form a cluster to which a particular meaning can eventually be attached (e.g., coffee and donuts)
- **Dictionary based on substitution:** a code is automatically attributed to a particular fragment if it contains a specified combination of words or phrases (e.g., code 'Food' is attributed if the fragment contains any of the following words: Bread, Tea, Broth, Breakfast, Lunch etc.)

# Dictionary based on substitution: an example

## **Human dignity (latent code)**

- Politeness
- Loyalty
- Pride
- Goodness
- Love
- Courage
- Male dignity
- Responsibility
- Understanding
- Decency

- Human rights
- Self-respect
- Will power
- Conscience
- Justice
- Rank
- Respect
- Intelligence
- Honour
- Honesty

The list was identified with the help of an open-ended question asked in an online survey

# Preconditions for quantitative content analysis

- **Parsing:** parts-of-speech (also known as POS, word classes, or syntactic categories) are useful because of the large amount of information they give about a word and its neighbors. Knowing whether a word is a noun or a verb tells us a lot about likely neighboring words (nouns are preceded by determiners and adjectives, verbs by nouns) and about the syntactic structure around the word (nouns are generally part of noun phrases), which makes part-of-speech tagging an important component of syntactic parsing
- **Tokenization:** segmenting running text into words and sentences
- Sentence tokenization methods work by building a binary classifier (based on a sequence of rules, or on machine learning) which decides if a period is part of the word or is a sentence boundary marker. In making this decision, it helps to know if the period is attached to a commonly used abbreviation; thus an abbreviation dictionary is useful

# Vectorization

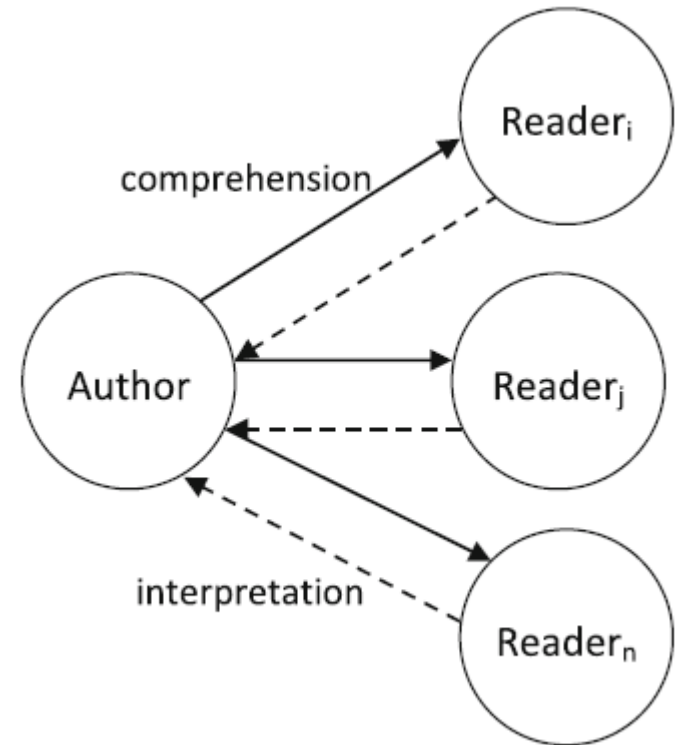
With the help of **tokenization**, the text is divided in sentences and then transformed into a table. These operations pave the way to **vectorization**. Sentences are rows in a table, variables (specific words or qualitative codes) are columns in the table

	Word <sub>1</sub> /Code <sub>1</sub>	Word <sub>2</sub> /Code <sub>2</sub>	...	Word <sub>i</sub> /Code <sub>i</sub>
Sentence <sub>1</sub>				
Sentence <sub>2</sub>				
Sentence <sub>3</sub>				
...				
Sentence <sub>i</sub>				

# Types of texts

**Rhetorical texts** (novel, poem, diary, or essay) have a loose structure. Metaphors and analogies abound in such texts. The content analysis of rhetorical texts calls for prioritizing **interpretation**: they aim at provoking free associations and creative thinking (cf. **subjectivist** approach)

**Stylistic texts** (scholarly article, textbook, or scientific letter) have a clear, often rigid structure. Arguments in stylistic texts must meet high logical standards: exhaustiveness and mutual exclusiveness of categories, transitivity and consistency in their rank ordering and so forth. **Comprehension** seems to be more appropriate for the content analysis of stylistic texts as a result of their orientation toward conveying a message in the least ambiguous manner (cf. **objectivist** approach)



# Triangulation: definition and forms

- ‘The combination of methodologies in the study of the same phenomenon’ (Jick, 1979)
- **Data** triangulation: the use of a variety of data sources
- **Investigator** triangulation: the use of more than one researcher
- **Theory** triangulation: using multiple perspective to interpret a single data set
- **Methodological** triangulation: the use of multiple research methods, including various forms of content analysis, to study one problem or document. **Mixed methods research** as practical manifestation

# Validity and reliability in content analysis

- Mixed methods content analysis [a form of investigator and methodological triangulation] allows assessing the reliability of coding
- A conventional approach to assessing the reliability of a content analysis is to calculate coefficients of inter-coder agreement: Krippendorff's alpha, Cohen's kappa, Bennett, Alpert and Goldstein's S and some others
- Mixed methods content analysis allows adding correlation coefficients (Pearson's  $r$  between the qualitative content analysis outcomes and the correlational analysis outcomes, for instance) and to this list. Cf. 'In content analysis [the use of correlation coefficients] is seriously misleading' (Krippendorff, 2004, p. 245)
- The basic table takes a different shape: coders are in its rows, codes – in its columns
- The choice of a proper measure depends on particularities of the text

# Choice of reliability measures and the type of text

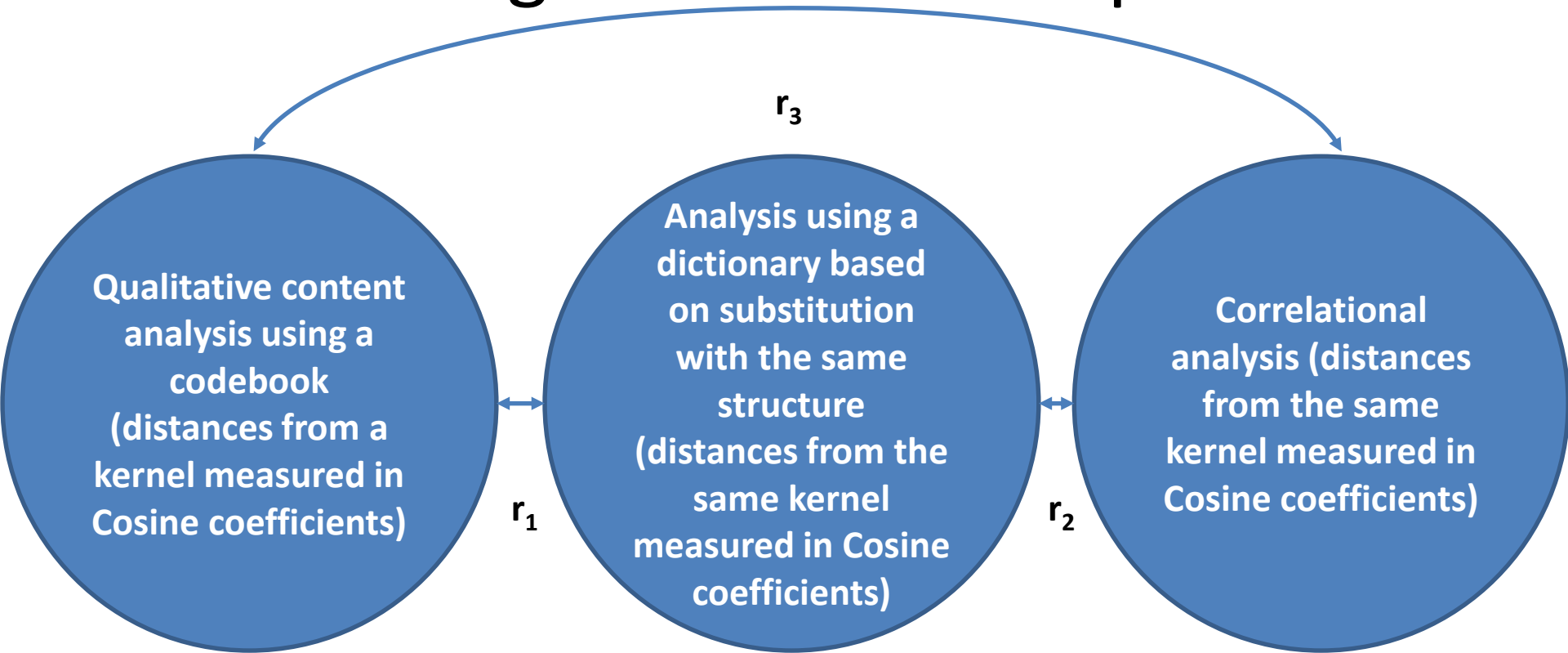
**Table 1** Reliability measures appropriate for particular types of text and research tasks

Type of reading	Text format	
	Stylistic	Rhetorical
Comprehension	Scott's $\pi$ , Krippendorff's $\alpha$ , Pearson's $r$ (between word co- interpretation occurrence and qualitative coding)	Perreault and Leigh's $I_r$
Interpretation	Perreault and Leigh's $I_r$	Bennett, Alpert and Goldstein's $S$ , Cohen's $\kappa$ , Pearson's $r$ (between dictionary based on substitution and qualitative coding)

Pearson's  $r$  is calculated for distances between vectors (i.e., documents) expressed as **Cosine coefficients**



# Using different types of content analysis when working on a dataset: a path-model



**Regression analysis** is also potentially possible, e.g., Results of the qualitative content analysis as a outcome variable (DV), Results of correlational analysis, Results of the analysis using a dictionary based on substitution and some controls (e.g., Time, Author's gender etc.) as predictor variables

# Computer programs for content analysis

- *QSR International*, Australia  
<http://www.qsrinternational.com/>: **NVivo**, **N6 (NUD\*IST)**, **XSight**
- *Provalis Research*, Canada (Montreal)  
<http://www.provalisresearch.com/>: **QDA Miner** (module for qualitative content analysis) and **WordStat** (module for the analysis of co-occurrences and the use of dictionaries based on substitution). These two modules used in conjunction open multiple opportunities for mixed methods research when content analyzing texts
- **ThinkMate**: an on-line platform that I am currently working on. It has some similar features with **QDA Miner**, but also offer some new options, such as the search of similarly minded people on the basis of reading (and coding) particular texts, i.e., a novel type of social network

# Qualitative coding using *QDA Miner*

The screenshot displays the QDA Miner software interface. The title bar reads "QDA Miner - Dignity\_sources\_coded.ppt". The menu bar includes "Project", "Cases", "Variables", "Codes", "Document", "Retrieval", "Analyze", and "Help".

On the left, there are two panels:

- CASES:** A list of cases from Case #1 to Case #17. Case #3 is selected.
- VARIABLES:** A list of variables. Under the "ARTICLE [DOCUMENT]" category, there are no variables listed.
- CODES:** A list of codes. Under the "Comparison" category, there are "Justification" and "Precedents". Under the "Practice" category, there are "Dialogue", "Elective\_affinity", and "Entrepreneurs".

The main window shows a document titled "ARTICLE". The text in the document is as follows:

{Bowman, James, 'The lost sense of honor', *The Public Interest*, Fall 2002, 32-49}

How difficult it often seems for our contemporaries to understand or use correctly the language of shame – and of honor, shame's complement and opposite 33.

The need for honor, particularly in a military context 35. Honor is incompatible with the spirit of our age. 8 reasons

1. Honor and shame are socially founded and not in the control of the individual who is honored or shamed
2. Honor is fundamentally elitist 36. The 'honor group'... is composed of one's equals even when differences of rank exist, since their opinion matters more – or ought to matter more – than that of any outsider
3. Honor is judgmental 37.
4. Uncompassionate: There can be no trust where there is not also the possibility of a breach of trust, and vice versa. In the same way, there can be no honor where there is not the possibility of shame and disgrace 38.
5. Honor is a relativistic standard. Honor... depends on context 38. Varies from society to society
6. Honor has been at odds with Christianity 39. No second chest
8. Honor... is fundamentally different for men and women 41. Bravery vs. chastity

WWI: the individual acts of bravery and heroism on which honor depends had been rendered meaningless 43.

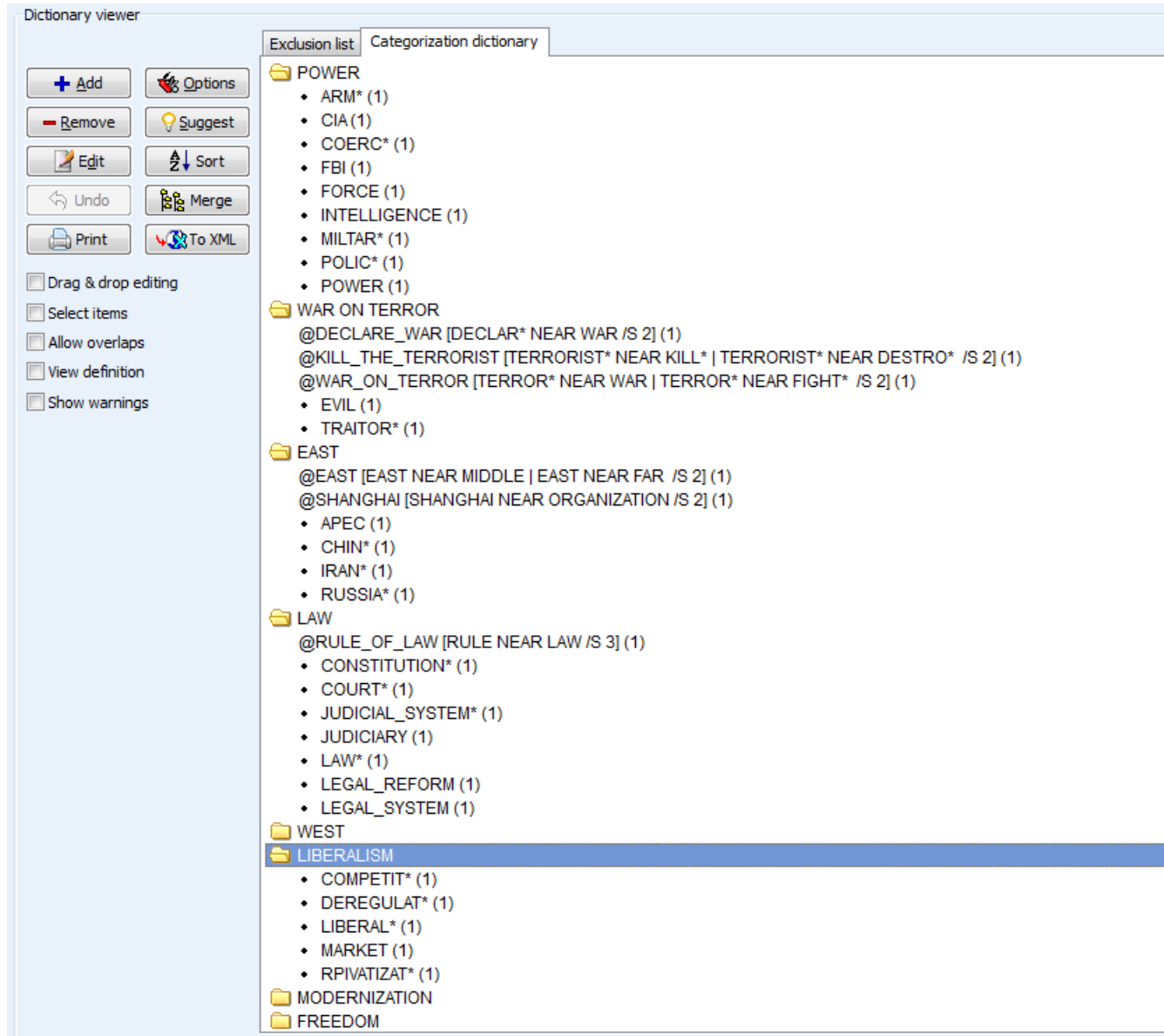
Once claims of the individual were recognized as paramount, they of necessity brought into disrepute and scorn the claims of the group [including honor] 44.

Islam: The old imperatives of honor that we in the West have almost forgotten 46.

Arab culture: [the] concern is solely with the pride and honor of not showing fear – and therefore, in an extreme honor culture, of insisting that one loves what other fears, namely death 47. An unreformed honor culture

On the right side of the main window, there is a vertical axis with labels: "Honor", "Relativism", and "Honor".

# Dictionary based on substitution in *WordStat*



# 25 most frequent words used by Lord Byron

## (analysis of co-occurrences using *WordStat*)

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
LOVE	891	0.80%	0.60%	0.30%	115	46.40%	297.4
EYE	719	0.60%	0.50%	0.20%	98	39.50%	289.9
MAN	718	0.60%	0.50%	0.20%	65	26.20%	417.5
HEART	668	0.60%	0.50%	0.20%	117	47.20%	217.9
DAY	529	0.50%	0.40%	0.20%	87	35.10%	240.7
TIME	461	0.40%	0.30%	0.10%	71	28.60%	250.4
LONG	460	0.40%	0.30%	0.10%	70	28.20%	252.7
LIFE	438	0.40%	0.30%	0.10%	66	26.60%	251.8
THING	422	0.40%	0.30%	0.10%	36	14.50%	353.7
SOUL	379	0.30%	0.30%	0.10%	91	36.70%	165
MAKE	368	0.30%	0.30%	0.10%	43	17.30%	280
JUAN	366	0.30%	0.30%	0.10%	2	0.80%	766.2
LEAVE	365	0.30%	0.30%	0.10%	57	23.00%	233.1
HAND	359	0.30%	0.30%	0.10%	50	20.20%	249.7
FALL	330	0.30%	0.20%	0.10%	58	23.40%	208.2
GREAT	325	0.30%	0.20%	0.10%	33	13.30%	284.7
LIE	322	0.30%	0.20%	0.10%	61	24.60%	196.1
EARTH	317	0.30%	0.20%	0.10%	57	23.00%	202.4
DIE	314	0.30%	0.20%	0.10%	54	21.80%	207.9
FEEL	305	0.30%	0.20%	0.10%	60	24.20%	188
HOURL	305	0.30%	0.20%	0.10%	73	29.40%	162
HIGH	302	0.30%	0.20%	0.10%	44	17.70%	226.8
HATH	293	0.30%	0.20%	0.10%	44	17.70%	220
HEAR	291	0.30%	0.20%	0.10%	55	22.20%	190.3

# ThinkMate (beta-version)

The screenshot shows a web browser window with the address bar displaying 'social.r.smglab.ru/en/'. The browser's toolbar includes various icons for navigation and utility. Below the browser window, the ThinkMate website is visible. The header features the 'Thinkmate' logo on the left and a user profile 'Anton' on the right. A dark sidebar on the left contains a 'MENU' section with links to 'Projects', 'Индивидуальные книги кодов', 'О системе', and 'Contact us'. The main content area has a light gray background and features the 'Thinkmate' logo in blue. Below the logo, there is a paragraph of text explaining the platform's purpose: 'You have read an article and would like to have the notes in its margins at your disposal in the future? You have highlighted an element of a picture and do not want to lose it? You are interested in comparing your thoughts with respect to this article or picture with the other people looking at them? Thinkmate gives you a chance to accomplish these and many other tasks. The relevant data will be available to you wherever you are: at home, in the office, on the go.' This is followed by another paragraph discussing qualitative data and content analysis, stating that the platform allows users to store ideas, find similar-minded people, and create groups based on shared worldviews.

social.r.smglab.ru/en/

MUN mail Sign In Gmail Google.Ru Google.Ca The Globe and Mail Українська правда Gasera.Ru Compromat.Ru VKontakte Online PDF Converter PDFSplit NEWSru.com Date calculator Online Courses

Thinkmate

Anton

MENU

- Projects
- Индивидуальные книги кодов
- О системе
- Contact us

## Thinkmate

You have read an article and would like to have the notes in its margins at your disposal in the future? You have highlighted an element of a picture and do not want to lose it? You are interested in comparing your thoughts with respect to this article or picture with the other people looking at them? *Thinkmate* gives you a chance to accomplish these and many other tasks. The relevant data will be available to you wherever you are: at home, in the office, on the go.

Qualitative data (texts, visual images) are notoriously difficult to compress, aggregate, store and manage. Content analysis (both qualitative, i.e. manual coding of texts and images, and quantitative, i.e. the analysis of word co-occurrences) offers a possible solution. With *Thinkmate*, you will be able to identify the most interesting – from your point of view – fragments of a text/image and label (code) them in a particular manner. This on-line platform allows you to store the ideas that emerge in the process of reading a text or looking at an image. However, what the platform potentially offers goes well beyond that: it will help you to find similarly minded people by comparing their ideas and yours (codebooks and coded fragments). It might well become a new type of social network. Instead of common friends, place of residence, education etc., *Thinkmate* uses common ideas and a shared worldview as a basis for creating groups.

# http://social.r.smglab.ru/project/view/120/

Thinkmate

☰

МЕНЮ

Projects

Индивидуальные книги кодов

О системе

Contact us

Проект

Документ

Поиск согласия

Свойства

Документ: Spbnews\_Facebook.docx

Кодирование

Комментирование

Выберите код кодир

Отметить

Добавить код

Spbnews\_Facebook.c

Перейти

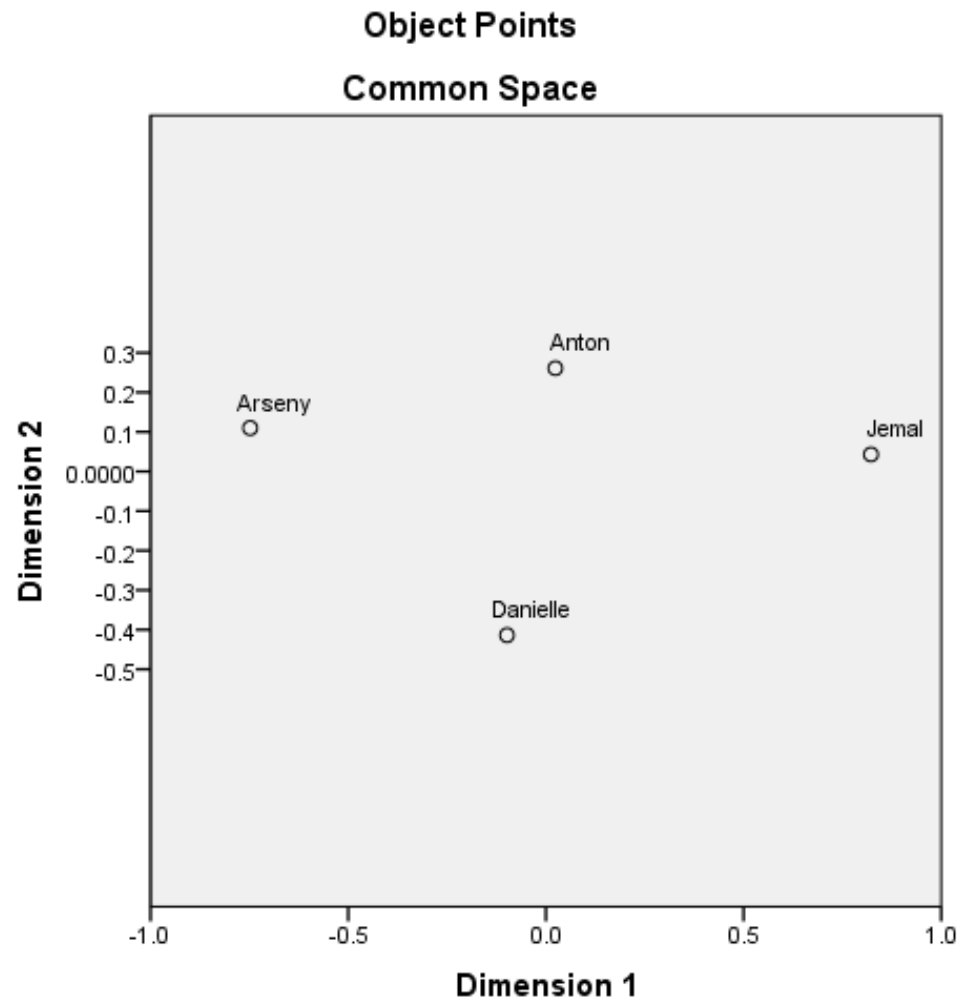
Скачать документ

Экспорт документа

Документ: Spbnews\_Facebook.docx

House Democrats released Thursday that were purchased by s Internet Research Agency to interfere with US politics. Representative Adam Schiff (D-California), the ranking member of the House Intelligence Committee, said that the release was meant to protect the U. S. from similar interference going forward. Sunlight is always the best disinfectant, Schiff said in a tweet. The 3519 ads released by the Democratic members of the House Intelligence Committee were purchased by the Internet Research Agency between early 2015 and late 2017. Some of the ads voice support for political candidates including President Trump as well as Bernie Sanders, while others oppose Trump, or support causes like Black Lives Matter and LGBT rights, sought to divide us by our race, our country of origin, our religion, and our politics, said Schiff. They attempted to hijack legitimate events meant to do good teaching self-defense, providing legal aid as well as those events meant to widen a rift. The link between these ads and Russia s Internet Research Agency was first reported last fall, when acknowledged that the agency had in its attempts to boost its messages on the social network. Democrats had last November, also told lawmakers last year that more than 11. 4 million people in the US had been exposed to these ads. However, Russia s disinformation campaign didn t stop with paid content. Facebook pages linked to the Internet Research Agency also created around 80,000 regular posts, which are estimated to have reached more than 126 million Americans. House Democrats said Thursday that they planned to release these 80,000 posts in the future as well. In a response to Thursday s release, Facebook published a that reiterated some of the steps the company is talking to prevent such abuse in the future. However, the post also acknowledged that the company could not completely rule out future disinformation campaigns: This will never be a solved problem because we re up against determined, creative and well-funded adversaries. But we are making steady progress.

# Map of close-mindedness





# Further development of *ThinkMate* (possible CS-4750 course projects)

- The **tokenization** algorithm needs improving. As of today, the algorithm is rather simplistic:
  - «dot + space + uppercase letter» OR «exclamation mark + space + uppercase letter» OR «question mark + space + uppercase letter»
- The content analysis of **images** could be done along the same lines, but relevant algorithms have to be adapted and adjusted accordingly
- The **multidimensional scaling** is not incorporated yet. Without the MDS one cannot really identify his or her similarly minded fellows
- The algorithm for **comparing the structure of the codebooks** also needs developing and so forth

# Additional sources

- Oleinik, A., 'What neural networks can't do? On artificial creativity' // *Big Data & Society* (under review)
- Oleinik, A., Popova, I., Kirdina, S., Shatalova T. (2016), 'On academic reading: citation patterns and beyond' // *Scientometrics*, 113(1), 417-435
- Oleinik, A. (2015), 'The language of power: a content analysis of presidential addresses in North America and the Former Soviet Union, 1993–2012' // *International Journal of the Sociology of Language*, 236, 181–204
- Oleinik, A. (2015), 'On content analysis of images of mass protests: a case of data triangulation' // *Quality & Quantity*, 49(5), 2203-220
- Oleinik, A., Popova, I., Kirdina, S., Shatalova T. (2014), 'On the choice of measures of reliability and validity in the content-analysis of texts' // *Quality & Quantity*, 48(5), 2703-2718
- Oleinik, A. (2011), 'Mixing quantitative and qualitative content analysis: triangulation at work' // *Quality & Quantity*, 45(4), 859-873