Malwina Kotowicz
Department of Computer Science & Media Technology
Linnaeus University
mk224ih@student.lnu.se

Introduction

The purpose of this document is to present and discuss ideas (i.e. chosen problems and solutions to them) for the **final project** within the course Programming for Digital Humanities. Therefore, for the sake of simplicity, this document was divided into following parts: **1) Problem statement, 2) Methodology, 3) Major blocks of code, 4) Expected outcome**.

Problem statement

Thematic Analysis is a method of identifying and analyzing patterns of meanings in a dataset. It illustrates themes that are important in description of the phenomenon under study (Joffe, 2012). As one of the most common methods within the field of qualitative analysis, it's widely used with data from, for example: field research, interviews, observation, solicited diaries and story completion (Braun and Clarke, 2006). This method consists of few phases, one of which is *coding* (tagging items of interest in data with a label: a few words or a short phrase). Such *content analysis* helps to primarily identify *themes* in next phase.

Thematic Analysis is widely used in psychology and social sciences; among many, examples include: studies on children's and parents' views on interventions to promote healthy lifestyle in primary schools (McCullogh et al., 2019), searching for thematic patterns in children's writing (DuCharme, 1994) or thematic content analysis of children's food advertising (Roberts and Pettigrew, 2007).

In this final project, I propose the design and implementation of an auxiliary tool for thematic analysis. The preliminary purpose of such prototype would be to serve child psychologists and counselors in initial assessment of child welfare. *Behavior Diaries* (solicited diaries) and story completion are among main qualitative research methods and are commonly used in developmental psychology (Gott, 2015; Lämsä et al, 2012). Children are often asked to journal about their everyday life or given topic. The tool I propose could be deployed as next step in analysis of such text. Moreover, it would accelerate analysis at early stage when the case study group consist of many participants (e.g. in metastudies on members of one U.S. state).

Additionally, such tool can be used to perform content analysis on any given text (be it transcribed interviews, customers' opinions, tweets or researcher's notes on a subject).

Methodology

Tasks to analyze datasets of big computational complexity and volume are conceptually challenging and often only possible with machines' computing capacities rather than humans alone. Hence, implementation of programming solutions is justifiable and facilitates this task significantly.

Because of datasets used in studies on *behavior diaries* are confidential and available only to researchers and counselors, I've chosen three children's blogs (due to similarity of blogs to journals and diaries) of high originality and distinct leading topics to test the tool on: jakes-bones, lidbemchild, neverisconds.

Programming approaches used in this project would consist of: 1) acquiring the content of several dozens of posts (e.g. from distinct periods in blog's history) of each blog, 2) Text *prettifying* (removing all metadata and clearing text content) and processing (extracting information of interest), 3) Sentiment analysis, 4) Visualization of data obtained in previous steps and, additionally: 5) Training a simple Machine Learning model on textual data extracted from blogs and 6) Preforming Topic Modeling with the model.

Major blocks of code

This project consists of following required blocks of code:

1) Web scraping and parsing performed with **BeautifulSoup**

2) Text preprocessing with **RE** (Regular Expressions: to identify patterns in text and clean undesirable metadata), **NLTK** (Natural Language Toolkit: a platform to work with human language data: for further text processing) and **TextBlob** (Natural Language Processing library for sentiment analysis: extracting polarity and subjectivity from text content)

3) Data visualization with **Plotly, Plotly Express** (Sentiment visualizations with line chart, topic tracking: visualization of ten most common words in frequency distribution of all words with bar chart and tag cloud, the longest sentences visualization: heatmap with **Seaborn**) and **Dash** (wrapping Plotly-based dashboard with Dash framework for interactive user interface)

4) Machine Learning model training with **Word2Vec** and **Gensim** (model used to learn vector representations of words: *words embedding*. Here the model will be trained on real data extracted from blogs)

5) Topic modeling with **LDA** model (allows to identify topics that best describe given blog sample and map them to fixed set of topics–sets of words in LDA model)

Expected outcome

The tool proposed in this project will process, analyze and visualize data in a manner described in previous section (*Major blocks of code*). Therefore, several outputs (outcomes) will be produced:

1) After textual processing of data with **NLP** tools, content of each blog (several dozens of posts per blog) will be visualized with **Plotly** and **Plotly Express**. Each blog will therefore *get* its own interactive dashboard of visualization (implemented with **Dash**) of sentiment analysis (line chart), the most frequently used words (bar chart and tag cloud) and the longest sentences (heatmap with **Python** and **Seaborn**)

2) Machine Learning model trained on blogs' data: **Word2Vec** model. Additionally, model training will after all allow some interesting features, for example exploring synonyms (function *.most_similar*) learned only from blog's content or mathematical operations *on words* (possible thanks to vector representations of words, e.g. *'horrible'* + *'awful'* = *'terrible'*)

3) Visualization of topic modelling with **LDA** (bubble graphs with interactive features). Topic modelling could help with thematic codes' designing, it will also illustrate tendencies, attitudes and dominating moods in children's writing.

4) GitHub repository of the project, uploaded here.

REFERENCES

1. DuCharme, C. C. (1994). Thematic patterns in children's writing: Human behavior evoking personal meaning. Day Care & Early Education, 22(2), 4–11

2. Joffe H. Thematic analysis. In: Harper D, Thompson AR, editors. Qualitative research methods in mental health and psychotherapy. Chichester: John Wiley & Sons, Ltd.; 2012. p. 209–223.

3. McCullogh N, Boyle SE, Fothergill M, Defeyter MA (2019) 'A really good balance': Thematic analysis of stakeholders' views on classroom- and games-based positive choices interventions for primary school children. PLoS ONE 14(7)

4. Michele Roberts & Simone Pettigrew (2007) A thematic content analysis of children's food advertising, International Journal of Advertising, 26:3, 357-367

5. Sue Gott (2015). Behaviour Diaries: An Assessment Tool for Supporting Children with Behavioural Difficulties, Speechmark Publishing

6. Tiina Lämsä, Anna Rönkä, Pirjo-Liisa Poikonen & Kaisa Malinen (2012) The child diary as a research tool, Early Child Development and Care, 182:3-4, 469-486

7. Virginia Braun & Victoria Clarke (2006) Using thematic analysis in psychology, Qualitative Research in Psychology, 3:2, 77-101

Web-based material, Linnaeus University and others:

Blogs:

1. Libdem child: http://libdemchild.blogspot.com/
2. Jake’s Bones: http://www.jakes-bones.com/
3. NeverSeconds: http://neverseconds.blogspot.com/

Other resources:

1. Word2Vec: https://medium.com/explore-artificial-intelligence/word2vec-a-baby-step-in-deep-learning-but-a-giant-leap-towards-natural-language-processing-40fe4e8602ba
2. LDA topic modelling: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158
3. DataWorkshop tutorials for ML challenge participants: https://dataworkshop.eu/en/
4. Sentiment analysis tutorials: https://towardsdatascience.com/real-time-twitter-sentiment-analysis-for-brand-improvement-and-topic-tracking-chapter-2-3-1caf05346721