Sentiment analysis of web information

Team name: eMotionLess

Vision of the project results

The finished project will be a command line-based app, capable of analysing a given piece of text in Lithuanian and determining the obvious and the hidden emotions of the text. The application will take a raw text or an URL to a news article as an input, scraping the website for the article, then analysing the text, looking at how present are each of the six main emotions (sadness, happiness, fear, anger, surprise, disgust) and then presenting the results in percentages.

Preliminary analysis of the problem

News articles usually contain some emotions that are not explicitly shown and there is a need for a tool that can detect such emotions in a text.

Functional requirements

The application should be:

- console based and have no extra GUI, runnable on UNIX OS,
- able to determine if input was appropriate (actual text, correct language etc.), whether it was raw text or an URL to supported news site,
- able to scrape main Lithuanian news sites (like delfi.lt or lrt.lt) for the headline and article,
- not too difficult to add support for other languages/sites.

Meeting Notes

Meeting notes 09-08:

Decided on team name: eMotionLess

Project: Focus on one language first, make a hand-made data.

Later split the work so someone is responsible for generating data sets (from real websites) and others are responsible for sentiment analysis.

Things done during the meeting:

- Time tracking was set up (manual entries to online excel sheet).
- GitLab repository was set up.

For next time: Each member should make a draft document that includes

1. vision of the project results,
2. preliminary analysis of the problem, functional requirements,

3. meeting notes (group meeting),
4. and a list of possible questions for the meeting with the client/supervisor.

Next meeting: Monday morning before lectures.

Meeting notes 09-11:

Vision of the project result: Anupras was closest to what we want. We just needed to mention six main emotions.

Preliminary analysis of the problem, functional requirements: Anupras again.

Questions for the client/supervisor we had:

- Are we free to choose websites it supports?
- Should we have forbidden articles (censorship)?
- Can we use data scraping tools or should we make our own?
- Should we go with machine learning or word based approach?

Extras:

- News sites: Compare Delfi to LRT. Privately owned vs national funded (More emotional vs less?)
- Preliminary roles: Lead programmer - Anupras, Note taker / coordinator - Domas.

Next meeting: Review of the final draft. Meeting Wednesday after lectures.

Meeting notes 09-13:

We overviewed Anupras's final draft and went through it to make it our final document.

Next meeting is Friday morning (weekly). Nothing yet planned for it.

Questions for supervisor

- Are we free to choose whether we go with a keyword based approach or machine-learning based approach?

Questions for the client

- How should the results be represented, i.e. should the percentages add up to 100% and show the distribution of emotions, or show how likely the text is to be in that emotion?
- Are six main emotions (sadness, happiness, fear, anger, surprise, disgust) all right, should there be more/less, should they be different?
- What news websites should it support?