Project Plan
Data<>Art Methods
December 2021
Inge Kengen

-- Concept

First; the concept of this project plan is about my own personal project. This project I'm starting just now, is using the Indonesian Language (also known as Bahasa Indonesia) as the input. Personally I am connected to this language via my Indonesian background that I inherited through my mom who's the first generation who set foot to Dutch land. Because of that, I had many indirect ways in how Indonesian influences came to me.

Up until now the Indonesian language is still a big question, in which I hope to have one day an answer to. For now, I want to use the language which I only feel familiar with, but do not understand, in my personal project.

Since it's still in development I'm not quite sure where I'd like to end up, but the essence of the work has to do with the Online way of how Indonesian communicate under a post on facebook. A big part of my friends are family members, which half of those, are Indonesian. I'm in a group that is mostly people from the village where my mom grew up, Sepang the village is. I'm in this group so I would be a little bit more updated to what happens, and what they talk about. I still don't get what they're saying, even tho I can follow a person-to-person conversation. But the text is build up out of many abbreviations, which is making me so confused.

To give an example:

> "Kl memang peduli sm kmpg harusnya di bangun infrastrukturnya....jgn pd saat kmpanye sj krn mau di dukung br koar2 sanasini peduli sepang .....dari sy lahir smpe skrg jalanan msh bgtu2 sj .....mgkn kl di tenggelamkan di buatkan jalanan yg lbh bagus ."

How I read it:

> "Kl **(Kl what?)** memang peduli sm **(sm in what way...?)** kmpg **(something with kilometers?)** harusnya di bangun infrastrukturnya **(infrastructure!)** jgn pd saat kmpanye sj krn mau**(want)** di dukung br koar2 **(2. what is this 2?!**) sanasini peduli sepang **(the village!)**.....dari sy lahir smpe skrg jalanan**(walking)** msh bgtu2**(...again the 2?)** sj .....mgkn kl di tenggelamkan di buatkan jalanan**(walking)** yg lbh bagus**(back)** ."

-- Production plan

I want to use the confusion I described in the concept, and make a project that gives you little clearance, but just enough, and that it keeps you engaged. The initial idea (in which I did a little demonstration 4 days after this Art-data course) was to gather comments from Facebook (like the one above) look up what the words could mean and print all the findings out, and creating new sentences. This could be the real sentence, or a not proper sentence.

Since I only had 4 days, in which I had to work 2 of them, I only had some evenings to make it all work..

For this I needed the following for at least some output:

- Comments from facebook
- Cutter to cut words to the "base" of one word. (Example: talked = talk)
- Online Indonesian Dictionary
- Online Indonesian Abbreviations Dictionary
- Connection to printer
- Printing words
- Making sentences
- Me reading the comments to the "audience" (my classmates)
- Take feedback for next steps

It was not a realistic plan, to have done all these steps in a few evenings. But; I DID get the following:

- ~~Comments from facebook~~
- Cutter to cut words to the "base" of one word. (Example: talked = talk)
- ~~Online Indonesian Dictionary~~, but not used yet with code, but by hand...
- ~~Online Indonesian Abbreviations Dictionary~~ But not used yet with code, but by hand...
- Connection to printer a TSP ticket printer
- Printing words (by images, I cheated...)
- ~~Making sentences~~
- Me reading the comments to the "audience" (My classmates)
- Take feedback for next steps

For the future next steps I'd like to gather the following:

- Live comments on facebook

>> Which will be going through a >>

- Cutter to cut words to the "base" of one word. (Example: talked = talk)

>> These words are going through the >>

- Online Indonesian Abbreviations Dictionary, to see if the word is a known abbreviation;

> If not; then send through the >>

- Online Indonesian Dictionary, to find if a word is a word on itself already;

> If not;

  - Look up the first letter of the word
  - Look up last letter of the word

- Print out found words on a TSP ticket printer
- Make a new "full" sentence with words found

-- Output

Since I know what I'd like to gather, I have many options in how to show it; of course I need to find the "best one".

All possible formats that could work for me:

1. A performance in which I create sentences with words that are printed from 1 comment, and make a sentence with it that could be. Do this a numerous amount of times.
2. A setup (installation) in where all tickets with words keep on coming out until one ticket is printed with the best possible sentence. So it would store a database of all possible words, and gives you the sentences which could be.
3. A computer (installation) which is saying the words out loud, like a test, and gives you the end result in sound.
4. A book made out of comments that have to do with the main post. So it would be a feedback loop. If the comments are under an article, the comments should be able to tell you what the article is about. But because of the abbreviations; it can give you a new article. Book shows comments with all it's possible words that the program has searched, shows new comment and reads the next one. Eventually you figured out what it's about, and see the main post what the comments were about.
5. An interactive work, in which the viewer is correcting the comment with my program. So it would show you all possible words, in which you choose which one is correct, make a sentence because you choose, and correct the first comment without really correcting it.
6. Instead of interactive, I could also build a bot who'd do it automatically. Then it would show/read out loud, all the possibilities of words, and create a sentence.

Minimal format (I see this as a backup version that has to work, and is for the presentation):

2. A setup (installation) in where all tickets with words keep on coming out until one ticket is printed with the best possible sentence. So it would store a database of all possible words, and gives you the sentences which could be.
   (https://seotoolscentre.com/complex-sentence-generator)

Maximal format (I see this as my final work, which I'd like to present on a future exhibition; and show on website):

5. An interactive work, in which the viewer is correcting the comment with my program. So it would show you all possible words, in which you choose which one is correct, make a sentence because you choose, and correct the first comment without really correcting it.

-- Steps to take

Time estimation of the step: 2 weeks

1. I have many steps to take... First of all, I need to gather the **content** what I want to use for the whole project. Like which comments. This could be news/topics/posts, or really personal, like the one I used before; This was about the village where my mom was born, and how they should not sell their houses to big cooperations.
To scrape these comments of the Internet, I'd like to use **Python**, since this is the language which I find the nicest to work, and I can send stuff to other programs. (If I should need that).
So I already know that I have to install for scraping the following Python packages:
- requests, re, json, time, logging,
- collections, bs4 (BeautifulSoup)

Time estimation of the step: already ready

2. If I have the content, I need to **filter** it. It has to go first through a cutter/stemmer to get the stem of the word. For this I need to pip install pySastrawi and import SyemmerFactory. The only input it needs is the comment I want to get the "base" from.

Time estimation of the step: 3 days

3. With this "new" sentence I can look up the words in the dictionaries. I need to first make a **dictionary** of all the words so I can look up all the words separate from each other. This is just with python script. Then all the separate words will be a '**key**' and the search will give them all the found '**values**'.

Time estimation of the step: 3 weeks

4. All the keys will first go through the **abbreviation dictionary** (also known as Keteglo), I need to **find or build** this python script. But could use the same build up as the script that is using the KBBI (big dictionary bahasa Indonesia). (this in the next step) If the keys are not found in the Keteglo, then it will go into the KBBI. If they are found in the Keteglo, then it will put it in as a value to that key. If there's more possibilities found, then all will be added as a value.

Time estimation of the step: already done

5. For the big Indonesian dictionary, someone already made a python script, for this I need to install;
   - kbbi
   - requests
   - beautifulsoup4
   - appdirs

The only input I have to give to this is a word.

Time estimation of the step: 2 weeks

6. If the word is not recognized, I'd like to use a LCS (**Longest Common Subsequence**) in which it will give me the longest result that is possible out of the original word. I'm in contact with a person, (I showed you in the zoom-course Art-Data), who wrote a paper about it, and showed how this worked within Bahasa Indonesia. He could help me when he's done with his paper (after the 20st of December).

>>> See next steps on next page. >>>>>>>>>>>>>>>>

Time estimation of the step: 2 weeks

7. Before we go to step eight, I'm still in doubt if I want to use the **definition of the words**. It does make it less confusing, so the audience gets what the context is about. But since my main focus is the confusion I get from abbreviations, I'm not sure if I'd like to... If I want to use the definitions of the word, I need to **extend step 4 & 5**, and add as a value also the definition.

Time estimation of the step: 1 week

8. Since I want it to be **printed**, I need to connect this whole script to a printer. I bought a TSP100III, which is a ticket printer that I use at work. (for sending which drink/dish has to be made, and for which table). It's a printer with fast and quick results, which also cuts the print itself. So then I can have many words all over the floor. Lucky me, there's also a **python script** already for this; StarTSPImage. I need to figure out how to get it nicely printed, visually, for the 'audience' to see.

What is gonna be on this print then?
- The original word from the comment (point 0)
- The base word (after the stemming, if that is even an option)
- 1 optional word what the original could be
- Translations of this word

So if KL (from example in my concept) could mean:
Kalau/Kuala Lumpur/ Kurang Lancar

Output would be:

- KL
ticket 1. - kalau
- if

- KL
ticket 2. - Kuala Lumpur
- Kuala Lumpur

- KL
ticket 3. - Kurang Lancar
- Less smooth

Time estimation of the step: while busy

9. After all of this, I'd like to get a sentence as a result. But am not yet sure if I want it to be the best translation, or that I'm gonna play with it as well.

Time estimation needed in total:
10 weeks and 3 days

Counting from 20st of December. 2021 then it should be done on the 6th of march 2022, of course not really real, but that means it should be doable before the end of the year, which is great!
If I would count 5 weeks later from now, then at that point I have my semester 1 presentations. I hope by then I indeed came way further then where I came now, and figure out other steps necessary.

-- Proof of concept

For the presentation that followed up this course I had a small tryout; see the video I added next to this PDF (if you feel like watching of course). The presentation form was showing something without explaining your project/concept. So really what the 'audience' felt/saw/thought of it when they see it for the first time.
From that presentation I got the following feedback:

- felt like watching a secretary
- sound wise great, rhythm of talking over a print
- interesting how you're trying to keep up with the print
- constantly trying to guess which language it is
- the different lengths of messages are interesting
- do you actually know what you're saying? You once apologized because you said a word wrong. It seemed like it was the first time for you reading it too
- felt the opposite way, it was like you knew what everything was
- the setup on the desk is fitting, also that you asked the audience to stand on the other side
- would be nice if you'd have 1000s of receipts and have them pile up
- being able to anticipate the end made a big difference to the end

After explaining:

- you didn't sound very confused to the listener; where is the potential of making the viewer engaged and confused
- made me think of the "your unerasable text" by Stefan Tiefengraber where there is a print coming out of the printer, straight into a shredder.

-- Data resources

Data is all Online, this way, it feels like this project can be used as a tool. A tool which is fluid with any input it gets.

I'm using many build scripts already, so it's just a way of figuring out how to glue these together in the way I want.

I'm reaching out to the person who wrote the paper about micro text in Bahasa Indonesia, which is a big help already. Next to that a friend who's studying Data Science was interested in my project, and offered to help me by looking what I have and make it easier for myself.

I can reach out to Arthur Elsenaar who was my coach in previous years and helped me a lot with coding and what is all possible.

I'm in constant contact with Cocky Eek, who's my coach this year, and is helping me with being focused on my personal interest, The Essence of this work, the confusedness from Indonesian abbreviations.