**Arabic Tagging Project**

16th meeting – 22**.**0**9.**2020

**What we did**

In manage tags page we changed the way the user can open his option, now its right click on each terms and opens 6 options to add parent, child, edit name, change parent, delete, delete all. In addition we added a graph representation of the tags heirarchy, we want to connect it to the current representation of the tags in a treeview and when the user selects a term in the treeview it also gets selected in the graph, and then the user can edit it from both representation whichever he finds easier.

Users – the site not has 3 users, admin, user and guest. Guest is not yet implemented, user and admin are the only difference is who can add new users, for the next meeting the admin will have a page where he can add new users and guest.

**Ideas**

Add a graph representation of each poem where the user can see all terms and all their tags, make sure the tags are colored based on domain (root), place it in the tagging page (?) maybe add viz page and a button in tagging page that navigates to viz page with this poem and its graph.

**Confusion Matrix**

Given all poems the most occurred terms (top 5-10) are stop words, should we keep them? Yes.

We had a problem in choosing the size of the window we’re looking at, bigram or trigram.. and weather we look at lines or “half line”. The code that counts occurrences is ready we only need to defined the window.

Unigram matrix will be 220k X 220k, it will be דלילה since we only have 22,000 poems but we can get some תובנות from it and from there on move on to other form of analysis on them maybe try bigrams, if we want to find terms with משמעות דומה then we need to run clustering on the terms (word embedding or occurrences matrix).

Bigrams– run on the poem with window of 2 terms, get all possible options and check number of times each bigram occurs. This is a one time job that creates the initial data we need.

What will those bigrams give us?

Ali – (1) In a period we have something called formulas which are two terms that the writer uses them to describe something? (2) check which bigrams **don’t** come from the same domain but come together a lot (3) התנגדיות same line has for example big \_\_ \_\_ \_\_ small

Tsvika – bigrams we can do, we still need to check the abstract definition of the terms in order to check Ali’s (2).

Ilai – bigrams that are term1 term2 and term2 term1 are they the same? No, each one is its own bigram. Lets say we have a bigram that is not שכיח e.g. ענן במכנסיים this is a metaphor, but it is rare, meaning we might have the term ענן with many more bigrams with also higher freq. but they are not as interesting as the ones that are rare.

Does it matter which term is first and which is second in the bigram? Back to basics we need:חשוב לי לדעת השכיחות של צמדי מילים . this goes to sematic meaning of metaphores and weather it matters which term is first or second.

**Contradictions**

If we get an english database of terms and their contradictions and translate it, will that be enoght? Probably not.. different languages have different ways to represent contradictions and once transalted literaly a big part might not be correct.

What we can do? (1) Talk to Heyam, she has a list(lists) of contradictions in arabic that can help us as a basis. (2) web scrapper on Maani site and collect all contradictions from there.

**Things to do:**

* Tsviki – think about the clustering option and how we can make it better.
* Find all bigrams and their freq. and in which poems each bigram occures.
* Build graph for each poem, make sure its colored based on domain.
* Implement the whole login, signup…