Arabic Stemming

Agenda

- Text Normalization Tasks:
  - Stemming
  - Lemmatization
- Arabic Stemming and Stemmers.
- How to build Arabic stemmer.

```
>>> porter= nltk.PorterStemmer()
>>> stemming=[porter.stem(w) for w in tokens]
>>> for w in stemming:
        print w
```

حكى
بعضهم
قال
:
كنت
في
سفر
فضللت
عن
الطريق
،
فرأيت
بيتاً

```
>>> lancaster= nltk.LancasterStemmer()
>>> stemming=[lancaster.stem(w) for w in tokens]
>>> for w in stemming:
        print w + '-',
```

```
حكى- بعضهم- قال- :- كنت- في- سفر- فضللت- عن- الطريق- ،- فرأيت- بيتاً- في- الفلاة
. -فأتيته- فإذا- به- أعرابيّة- ،- فلما- رأتني- قالت- من- تكون- ؟- قلت- ضيف -،
قالت- أهلاً- ومرحباً- بالضيف- ،- انزل- على- الرحب- والسعة- .- قال- فنزلت- فقدمت
-لي- طعاماً- فأكلت- ،- وماءً- فشربت- ،- فبينما- أنا- على- ذلك- إذ- أقبل- صاحب
البيت- .- فقال- من- هذا- ؟- فقالت- ضيف- .- فقال- لا- أهلاً- ولا- مرحباً- ،- ما- لنا
-وللضيف- ،- فلما- سمعت- كلامه- ركبت- من- ساعتي- وسرت- ،- فلما- كان- من- الغد
رأيت- بيتاً- في- الفلاة- فقصدته- فإذا- فيه- أعرابيّة- فلما- رأتني- قالت- من- تكون
؟- قلت- ضيف- .- قالت- لا- أهلاً- ولا- مرحباً- بالضيف- ،- ما- لنا- وللضيف- ،- فبين
-ما- هي- تكلمني- إذ- أقبل- صاحب- البيت- فلما- رآني- قال- من- هذا- ؟- قالت- ضيف
.-
```

```
>>> wl=nltk.WordNetLemmatizer()
>>> rr=[wl.lemmatize(w) for w in tokens]
>>> for word in rr:
        print word + '-',
```

حكى- بعضهم- قال- :- كنت- في- سفر- فضللت- عن- الطريق- ،- فرأيت- بيتاً- في- الفلاة
. -فأتيته- فإذا- به- أعرابيّة- ،- فلما- رأتني- قالت- من- تكون- ؟- قلت- ضيف -، -
قالت- أهلاً- ومرحباً- بالضيف- ،- انزل- على- الرحب- والسعة- .- قال- فنزلت- فقدمت -
-لي- طعاماً- فأكلت- ،- وماءً- فشربت- ،- فبينما- أنا- على- ذلك- إذ- أقبل- صاحب -
البيت- .- فقال- من- هذا- ؟- فقالت- ضيف- .- فقال- لا- أهلاً- ولا- مرحباً- ،- ما- لنا
-وللضيف- ،- فلما- سمعت- كلامه- ركبت- من- ساعتي- وسرت- ،- فلما- كان- من- الغد -
رأيت- بيتاً- في- الفلاة- فقصدته- فإذا- فيه- أعرابيّة- فلما- رأتني- قالت- من- تكون
؟- قلت- ضيف- .- قالت- لا- أهلاً- ولا- مرحباً- بالضيف- ،- ما- لنا- وللضيف- ،- فبين -
-ما- هي- تكلمني- إذ- أقبل- صاحب- البيت- فلما- رآني- قال- من- هذا- ؟- قالت- ضيف
.-

Arabic Stemming

- What is stemming?

Stemming is the process of removing any affixes from words and reducing these words or return to their roots. For example, stemming the English word computing produces the root compute. This is the same root produced by the word computation.

- How is the stemmer implemented?

The first thing the stemmer does is remove the longest suffix and the longest prefix. It then matches the remaining word with the verbal and noun patterns, to extract the root.

Arabic Stemming

- What are stemming types?

1- Light Stemmer

Ex: طالب <=== الطالب

2- Root-Based Stemmer

Ex: الطالب ===> طلب

Arabic Stemmers

http://www.nltk.org/api/nltk.stem.html

- Assem's Arabic Light Stemmer (BETA)

https://arabicstemmer.com/

- ISRIStemmer.

  http://www.nltk.org/api/nltk.stem.html

- Tashaphyne Arabic Light Stemmer:

  https://pypi.python.org/pypi/Tashaphyne/

- Khoja Arabic Stemmer(root-based stemmer)

http://zeus.cs.pacificu.edu/shereen/research.htm

- FARASA stemmer

http://qatsdemo.cloudapp.net/farasa/register.html

Assem's Arabic Light Stemmer (BETA)

Description

Welcome to the Arabic Light Stemming Algorithm made for Snowball, it's fast and can be generated in many programming languages (through Snowball).

Demo

Type some Arabic text and press "**Stem!**" button or "**File**" to read from a local "**.txt**" file

مكتبة لمعالجة الكلمات العربية وتجذيعها

STEM! FILE

Stats words: 5 stems: 5 ratio: 1.2

مكتب معالج كلم عرب تجذيع

Please enter your text: **أدخل النص المراد معالجته:**

المدرسة
الطالب
يفكرون

Please note that there are some limitations to try the Dependency Parser:

- The demo is confined to process only three sentences per request. each sentence shouldn't exceed 20 words.
- The length of the text to be processed should be within 400 characters.

Lemmatization أصول الكلمات | Process text معالجة النص | Clear text مسح النص | **Text length** 21 عدد أحرف النص

مدرسة . طالب . فكر

```
5
6   word=u"الْعَرَبِيَّةُ"
7   ArListem=ArabicLightStemmer();
8   stem=ArListem.lightStem(word);
9   print stem
10
11
12  print ArListem.get_unvocalized();
13  word=u'أفتكاتبانني'
14  stem=ArListem.lightStem(word);
15  print ArListem.get_stem();
16  print ArListem.get_right();
17
18
19  ArListem=ArabicLightStemmer();
20  word=u'فتصبرين'
21  stem=ArListem.segment(word);
22  print str(ArListem.get_affix_list()).decode("unicode-escape");
23  print ArListem.get_segment_list();
24
```



```
Python 2.7.12 |Anaconda 4.1.1 (64-bit)| (default, Jun 29 2016, 11:07:13) [MSC v.1500 64 bit (AMD64)] on win32
In[2]: runfile('C:/Users/Amira/Anaconda2/Lib/site-packages/Tashaphyne/module2.py', wdir='C:/Users/Amira/Anaconda2/Lib/site-packages/Tasha
عراب
العرابية
كاتب
7
[{'prefix': u'ف', 'root': u'صبر', 'suffix': u'ين', 'stem': u'تصبر'}, {'prefix': u'فت', 'root': u'صبر', 'suffix': u'ين', 'stem': u'صبر'},
set([(1, 5), (2, 5), (0, 7)])
```

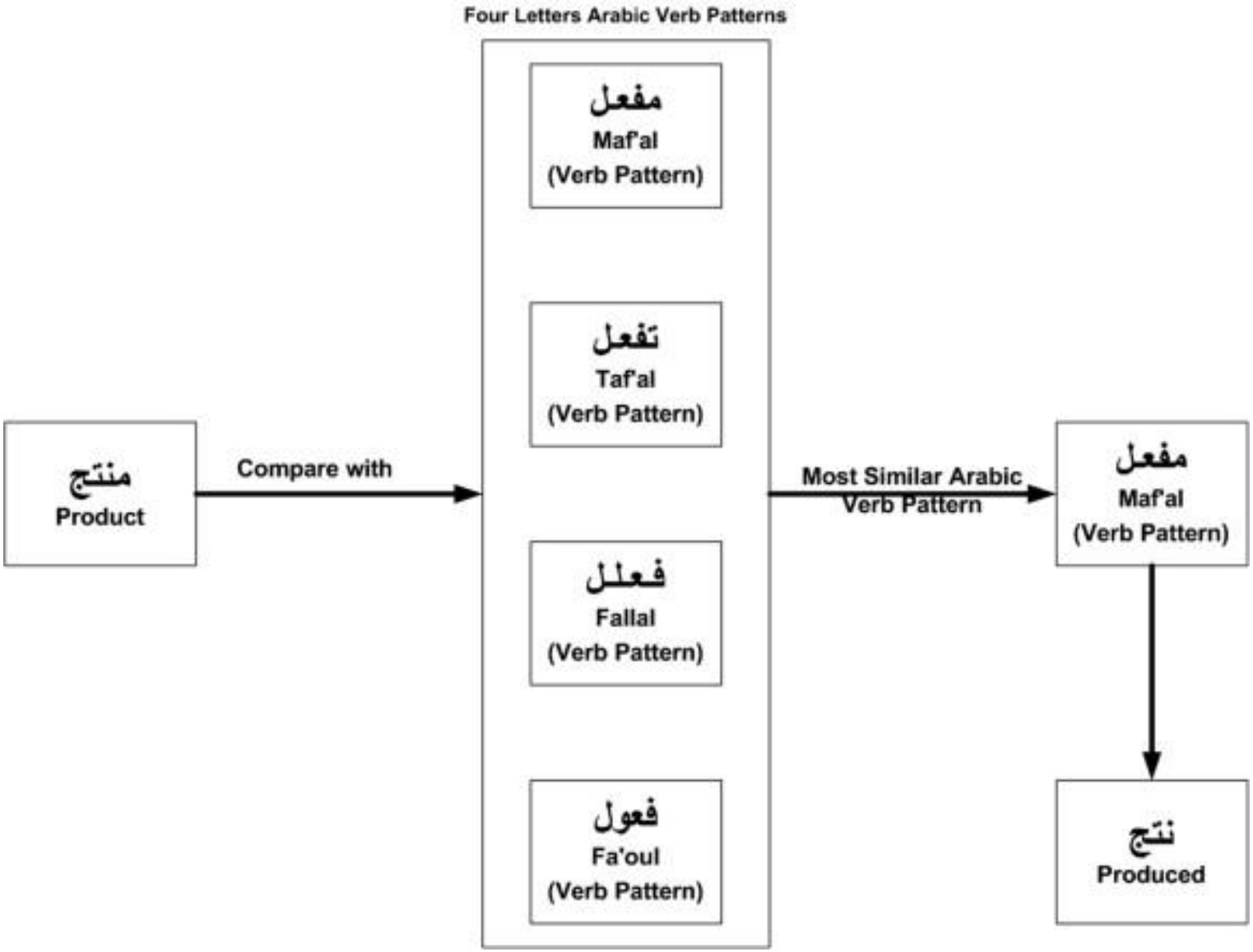
Arabic root-based stemmer algorithm

Input: A text file that contains the Arabic words

Output: Arabic Triliteral Verb/Verbs

1. Remove Arabic prefix(es) from each word
2. Normalize 3 shapes of (Alif, "أ إ آ") to (Bare Alif, "ا")
3. Remove suffix(es) from each word
4. Determine word length after removing affixes (prefixes and suffixes)
5. Identify Arabic patterns having same lengths to word length in step 4.
6. Compare each pattern identified in step 5 with extracted word from step 3
7. Select the closest pattern:
    a. Choose the pattern from the set of Arabic patterns having same lengths to word length which has the highest number of common Arabic letters with the Arabic word extracted from step 3.
    b. Determine the pattern which has the largest matching corresponding letters with the generated word from step 3 which is considered as the right pattern, where the corresponding Arabic letters within the extracted word from step 3 will not be compared with three Arabic letters (Faa', "ف"), (Ayn, "ع"), (Laam, "ل") within the pattern under consideration.
8. Eliminate all matched letters in step 7. The Arabic letters of the Arabic word extracted from step 3 which corresponds to the Arabic letters (Faa', "ف"), (Ayn, "ع"), and (Laam, "ل") in the selected pattern (found in step 7.a) are selected to constitute the extracted Arabic root.
9. Refine the extracted Arabic root by converting some of the Arabic letters.

Arabic root-based stemmer algorithm



Khoja stemmer algorithm

Algorithm: Khoja Root-Based Stemming Algorithm
Purpose: Stemming Arabic Words
Input:
- Dataset
- Stop-word list
- Assets and patterns files

Output: Stemmed Dataset
Procedure:
1. Replace initial إ، أ، آ with ا.
2. Stop-words removal.
3. Remove punctuation, non-letters and diacritics.
4. Remove definite articles from the beginning of the word.
5. Remove the letter (و) from the beginning of the word and (ة) from the end of the word.
6. Remove prefixes and suffixes
7. Comparing the resulting word to patterns stored in the dictionary, if the resulting root is meaningless the original word is returned without changes.

Figure 1- Khoja Stemmer Algorithm

ISRI Arabic Stemmer:

The Information Science Research Institute's (ISRI) Arabic stemmer shares many features with the Khoja stemmer. However, the main difference is that ISRI stemmer does not use root dictionary. Also, if a root is not found, ISRI stemmer returned normalized form, rather than returning the original unmodified word.

- Additional adjustments were made to improve the algorithm:

1- Adding 60 stop words.

2- Adding the pattern (تفاعيل) to ISRI pattern set.

3- The step 2 in the original algorithm was normalizing all hamza. This step is discarded because it increases the word ambiguities and changes the original root.

- The ISRI Stemmer requires that all tokens have Unicode string types.

ISRI Arabic Stemmer:

```
>>> st=nltk.ISRIStemmer()
>>> resultstem=[st.stem(w) for w in tokens]
>>> for w in resultstem:
        print w + '-',
```

حكى- بعض- قال- :- كنت- في- سفر- ضلل- عن- طرق- ،- فرأ- بيت- في- فلة- ،- فأت- فإ
-، -ذ- به- عرب- ،- فلم- رأت- قلت- من- تكون- ؟- قلت- ضيف- .- قلت- اهل- رحب- ضيف
-نزل- على- رحب- سعة- .- قال- نزل- قدم- لي- طعا- أكل- ،- وم،- شرب- ،- فبن- انا
-على- ذلك- اذ- قبل- صحب- بيت- .- فقل- من- هذا- ؟- فقل- ضيف- .- فقل- لا- اهل- ولا
رحب- ،- ما- لنا- ضيف- ،- فلم- سمع- كلم- ركب- من- سعت- وسر- ،- فلم- كان- من- لغ
د- رأت- بيت- في- فلة- قصد- فإذ- فيه- عرب- فلم- رأت- قلت- من- تكون- ؟- قلت- ضيف
قلت- لا- اهل- ولا- رحب- ضيف- ،- ما- لنا- ضيف- ،- فبن- هي- كلم- اذ- قبل- صحب -. -
-. -بيت- فلم- رآن- قال- من- هذا- ؟- قلت- ضيف -

```
>>> st= nltk.ISRIStemmer()
>>> st.stem('معلومات')
u'\u0639\u0644\u0645'
>>> print u'\u0639\u0644\u0645'
علم
```

```
>>> st= nltk.ISRIStemmer()
>>> st.stem('معلومات')
u'\u0639\u0644\u0645'
>>> print u'\u0639\u0644\u0645'
علم
```

Agenda

- Text Normalization Tasks:
  - Stemming
  - Lemmatization
- Arabic Stemming and Stemmers.
- How to build Arabic stemmer.

Arabic Root Word Resources:

Arabic Root Word Resources:

https://wahiduddin.net/words/arabic_glossary.htm

Examples of typical usage of the root are shown in parenthesis.

a-b	to be a father, ancestor, forefather (ab, abū)
'a-b-d	to serve, worship, be devoted to, show veneration ('abd, 'ibāda, ma'būd)
'a-d-l	to act justly, equitably or to make straight, set in order ('adl, a'dāla, ta'dīl)
'a-d-m	to be non-existent, disappeared, destroyed, devoid of ('adam, 'adīm)
'a-f-w	to be obliterated, effaced, eliminated (al-'afūw, 'afwīya, 'afā', isit'fā', 'āfin, mu'fan)
a-h-d	to unify, be one (al-ahad, ahadīya, uhādī)
a-kh-r	to postpone, defer, be last, final, ultimate (al-ākhir, ākhar, ukhrā, ta'khīr, mu'akhkhara)
a-l-h	to adore, deify, turn to another with intense feeling (ilāh, ilāhī)
'a-l-m	to know, have knowledge, be informed, teach, notify (al-'alīm, 'ilm, 'ilmiya, 'allam, u'lūma)
a-m-l	to hope, to look attentively, meditate, consider (amal, āmāl, āmil, muta'ammil)
a-n-s	to be familiar, friendly, sociable (uns, insī, ins, anīs)
'a-q-d	to tie, knit, make a knot, put together, join ('aqd, 'aqīda)
'a-q-l	to have the faculty of reasoning, comprehension ('aql, 'aqlī)

Arabic Root Word Resources:

Not secure https://studentsofquran.com/root-words/

A Directory of arabic root words and letters, singular and plurals from Juz 1 – Juz 30

Root Words | Singular/Plural | Blue Juz 1-30

Juz 1

Juz 2

Juz 3

Juz 4

Juz 5

Arabic Root Word Resources:

1-http://www.tyndalearchive.com/TABS/Lane/index.htm
2- https://wahiduddin.net/words/arabic_glossary.htm
3- https://studentsofquran.com/root-words/