Zeyrek

Release 0.1.0

Olga Bulat

May 22, 2020

CONTENTS

Zeyrek is a python morphological analyzer and lemmatizer for Turkish. It is a partial port of the Zemberek-NLP Tools (Morphology) Zeyrek can perform morphological analysis of Turkish text, returning all possible parses for each word, and lemmatize words, returning all possible base (non-inflected) forms of words.

CHAPTER

ONE

USAGE

To use Zeyrek, first create an instance of MorphAnalyzer class:

```
>>> import zeyrek
>>> analyzer = zeyrek.MorphAnalyzer()
```

Then, you can call its *analyze* method on words or texts to get all possible analyses:

```
>>> for parse in analyzer.analyze('benim')[0]:
...     print(parse)
Parse(word='benim', lemma='ben', pos='Noun', morphemes=['Noun', 'A3sg', 'P1sg'],␣
↪formatted='[ben:Noun] ben:Noun+A3sg+im:P1sg')
Parse(word='benim', lemma='ben', pos='Pron', morphemes=['Pron', 'A1sg', 'Gen'],␣
↪formatted='[ben:Pron,Pers] ben:Pron+A1sg+im:Gen')
Parse(word='benim', lemma='ben', pos='Verb', morphemes=['Noun', 'A3sg', 'Zero', 'Verb
↪', 'Pres', 'A1sg'], formatted='[ben:Noun] ben:Noun+A3sg|Zero→Verb+Pres+im:A1sg')
Parse(word='benim', lemma='ben', pos='Verb', morphemes=['Pron', 'A1sg', 'Zero', 'Verb
↪', 'Pres', 'A1sg'], formatted='[ben:Pron,Pers] ben:Pron+A1sg|Zero→Verb+Pres+im:A1sg
↪')
```

If you only need the base form of words, or lemmas, you can call *lemmatize*. It returns a list of tuples, with word itself and a list of possible lemmas:

```
>>> print(analyzer.lemmatize('benim'))
[('benim', ['ben'])]
```

CHAPTER

TWO

INSTALLATION

To install Zeyrek, run this command in your terminal:

```
$ pip install zeyrek
```

CHAPTER

THREE

USER GUIDE

Zeyrek's morphological analyzer returns instances of Parse object (based on pymorphy2's Parse), which is a wrapper of namedtuple class.

Parse object fields include:

- *word*: the word itself
- *lemma*: base form of the word, as found in a dictionary
- *pos*: part of speech of the word. Note: Turkish is an agglutinative language, which makes it quite different from widespread European languages. A word can usually be much longer, made of Inflection Groups (IG), which can correspond to words in other languages. Each of these IGs can have its own part of speech, and the part of speech of the word as a whole is determined by the part of speech of the last IG.
- *morphemes*: sequence of morphemes in the word, a list of strings - abbreviations of English names of morphemes.
- *formatted*: a human-readable string representation of the analysis. There are several kinds of possible formats. Default formatter shows the dictionary item and its part of speech, and morphemes (with their surfaces, if available), divided into inflectional groups by | character.

CHAPTER

FOUR

LICENSE

Licensed under MIT License. Zemberek, from parts of which Zeyrek was ported, is under Apache License, Version 2.0.

CHAPTER

FIVE

DISCLAIMER

This project is in alpha stage, so the API can change.