Osman Momoh

Project I: Demo

COMP 6751: Natural Language Analysis

Fall 2021

Purpose

This demo is to present some key results and limitations from the pipeline. You may also run the main function in order to see all results.

Tokenization

File ID: test/16213

Output:

zambia	's	kwacha	falls	at	weekly	auction	the	zambian
kwacha	fell	at	this	week's	foreign	exchange	auction	to	18.75
kwacha	to	the	dollar	from	last	week	's	16.95	,
the	bank	of	zambia	said	.	the	rate	was	the
lowest	since	the	auctions	resumed	two	weeks	ago	under	a
new	exchange	rate	system	worked	out	with	the	world	bank
and	international	monetary	fund	.	the	bank	of	zambia	said
it	received	370	bids	,	ranging	from	13.00	to	20.75
kwacha	,	for	the	six	mln	dlrs	on	offer	.
one	hundred	and	thirty	five	bids	were	successful	.	a
british	high	commission	spokesman	said	britain	would	put	eight	mln
stg	into	the	auction	at	a	rate	of	one	mln
a	week	as	soon	as	zambia	reached	a	full	agreement
with	the	imf	.	the	money	could	be	spent	only
on	goods	produced	and	supplied	by	british	firms	,	excluding
luxuries	and	defence	equipment	,	the	spokesman	added	.

The numbers are normalized by standard NLTK

We split dashed numbers into two tokens for number parsing

Sentence Splitting

File ID: training/267

Output:

1) INDONESIA UNLIKELY TO IMPORT PHILIPPINES COPRA
Indonesia is unlikely to import copra
from the Philippines in 1987 after importing 30,000 tonnes in
1986, the U.S. Embassy's annual agriculture report said.
2) The report said the 31 pct devaluation of the Indonesian
rupiah, an increase in import duties on copra and increases in
the price of Philippines copra have reduced the margin between
prices in the two countries.
3) Indonesia's copra production is forecast at 1.32 mln tonnes
in calendar 1987, up from 1.30 mln tonnes in 1986.

NLTK probably should have split the title into a new sentence

POS Tagging

File ID: training/267

Output:

('indonesia', 'NN') ('unlikely', 'JJ') ('to', 'TO') ('import', 'VB') ('philippines', 'NNS')
('copra', 'JJ') ('indonesia', 'NN') ('is', 'VBZ') ('unlikely', 'JJ') ('to', 'TO') ('import', 'VB')
('copra', 'NN') ('from', 'IN') ('the', 'DT') ('philippines', 'NNS') ('in', 'IN') ('1987', 'CD')
('after', 'IN') ('importing', 'VBG') ('30,000', 'CD') ('tonnes', 'NNS') ('in', 'IN') ('1986', 'CD')
(',', ',') ('the', 'DT') ('u.s.', 'JJ') ('embassy', 'NN') ("'s", 'POS') ('annual', 'JJ')
('agriculture', 'NN') ('report', 'NN') ('said', 'VBD') ('.', '.') ('the', 'DT') ('report', 'NN')
('said', 'VBD') ('the', 'DT') ('31', 'CD') ('pct', 'JJ') ('devaluation', 'NN') ('of', 'IN')
('the', 'DT') ('indonesian', 'JJ') ('rupiah', 'NN') (',', ',') ('an', 'DT') ('increase', 'NN')
('in', 'IN') ('import', 'JJ') ('duties', 'NNS') ('on', 'IN') ('copra', 'NN') ('and', 'CC')
('increases', 'NNS') ('in', 'IN') ('the', 'DT') ('price', 'NN') ('of', 'IN') ('philippines', 'NNS')
('copra', 'NNS') ('have', 'VBP') ('reduced', 'VBN') ('the', 'DT') ('margin', 'NN') ('between', 'IN')
('prices', 'NNS') ('in', 'IN') ('the', 'DT') ('two', 'CD') ('countries', 'NNS') ('.', '.')
('indonesia', 'NN') ("'s", 'POS') ('copra', 'NN') ('production', 'NN') ('is', 'VBZ') ('forecast', 'VBN')
('at', 'IN') ('1.32', 'CD') ('mln', 'NN') ('tonnes', 'NNS') ('in', 'IN') ('calendar', 'NN')
('1987', 'CD') (',', ',') ('up', 'RB') ('from', 'IN') ('1.30', 'CD') ('mln', 'NN')
('tonnes', 'NNS') ('in', 'IN') ('1986', 'CD') ('.', '.')

Number Normalization (and Parsing)

File ID: test/16213
Output:
(NUMBER (SMALLCARDINAL two))
(NUMBER (SMALLCARDINAL six))
(NUMBER (SMALLCARDINAL one))
(NUMBER
(SMALLCARDINAL one)
(LARGECARDINAL hundred)
(AND and)
(MEDIUMCARDINAL thirty))
(NUMBER
(SMALLCARDINAL one)
(LARGECARDINAL hundred)
(AND and)
(MEDIUMCARDINAL thirty)
(SMALLCARDINAL five))
(NUMBER (MEDIUMCARDINAL thirty))
(NUMBER (MEDIUMCARDINAL thirty) (SMALLCARDINAL five))
(NUMBER (SMALLCARDINAL five))
(NUMBER (SMALLCARDINAL eight))
(NUMBER (SMALLCARDINAL one))

We have parsed one hundred and thirty five. The intermediate parses are also shown, which is a limitation.

Date Parsing

File ID: training/4425
Output (partial):

(DATE (MONTH march))
(DATE (MONTH march) (DAY 15th))
(DATE (MONTH march))
(DATE (MONTH march) (DAY 28th))

File ID: test/15910
Output (partial):

(DATE (FULLYEAR 1986))
(DATE (FULLYEAR 1985))
(DATE (DAYDIG 04) (SEP /) (MONTHDIG 09) (SEP /) (YEAR 87))
(DATE (DAYDIG 03) (SEP /) (MONTHDIG 09) (SEP /) (YEAR 87))

The actual parser is much more powerful than the examples shown, however there were insufficient corpus examples to fully demonstrate it. Please see the DateParser.py CFG in source code for examples of dates that can be parsed (shown in comments).