NLP Internship Code Challenge

Abstract

Develop a classifier `name_classifier`, which checks whether a given string of text is a valid person name or not. Here, we suppose the string input is always ASCII characters. This doesn't mean you don't need to consider non-English person names. E.g. you need to correctly classify "Jun Wang" (Chinese name) as a valid person name.

* Objective of this challenge is to check your general knowledge/skills of NLP & ML.

Dataset you can use to train your classifier:

You can download the list of valid person names from dbpedia here

> -There are many other interesting dataset dbpedia provides, which can be useful in the challenge. You are free to download them from here, and use them to improve your classifier.

List of common English words can be found here

> -This is useful for getting samples of strings which are not valid person names

Note that your classifier needs to be able to work on names (or non-name strings) which never appear on the dataset provided above, and will form part of our evaluation of your code.

You're free to use any dataset/dictionary from Internet, feel free to form your own dictionaries.

Example of Names/Strings Which Need to Be Classified

String	Label (True if name, else False)	Note
Jun Wang	True	Chinese name
Preembarrass Hippogryph	False	2 random words combined
Fustellatrici Pazze Perugia e dintorni	False	Some random Italian phrase from the web
Nishant Dahad	True	Indian name
Alison Cheung	True	English first name + Chinese surname
Undercloth Reclothe	False	2 random words combined
Chinese New Year	False	Proper noun of an event

Thames River	False	Proper noun of a river
Naomi Nguyen	True	Japanese first name + Vietnamese surname

Libraries

You are allowed to use any standard Python libraries. Except standard libraries, you’re allowed to use following Machine Learning & NLP related libraries.

scikit-learn

numpy

scipy

matplotlib

nltk

pandas

gensim

TensorFlow

Theano

Pylearn2

Pattern

MITIE

Unidecode

polyglot

Report

Provide a report about performance of your classifier together with your code. Please include precision, recall, f1, auc scores together with examples of misclassified strings.