Stack Overflow Tag Predictor

Classifying Stack Overflow Posts with NLP

Quinn Dizon, Mindy Zhou

Data & Packages

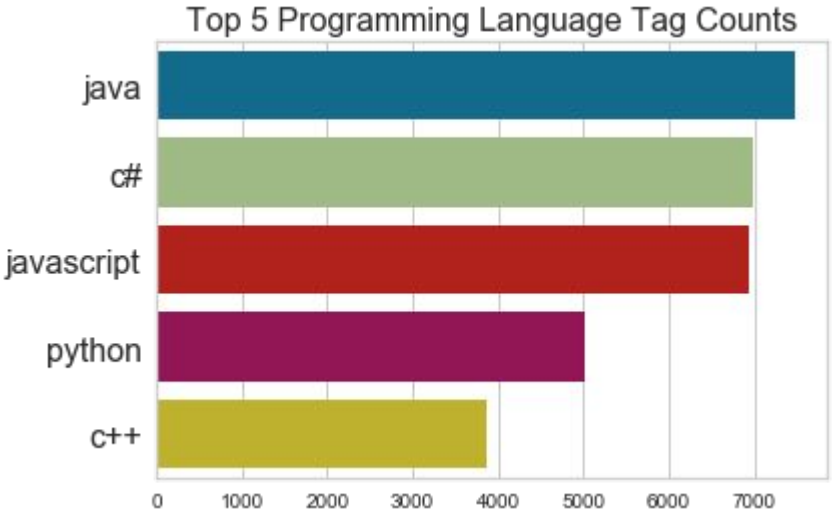
Dataset:

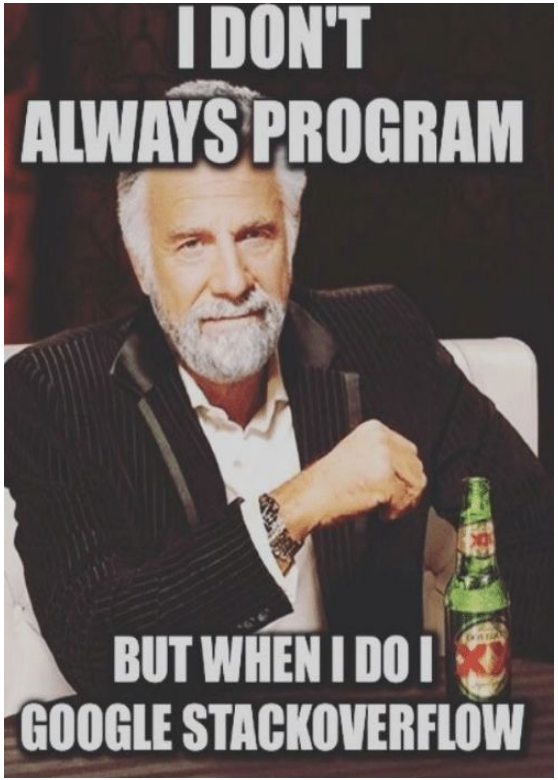
- 30,000+ raw posts from Stack Overflow
  - **Post text** - input strings
  - **Tags** - Classification labels

Packages:

- **NLTK** - text preprocessing
- **Sklearn** - machine learning
- **Gensim** - topic clustering
- **Yellowbrick** - evaluation visuals

Top Tags (languages)





Text Preprocessing

1. Input

How do I install a new package with pip?

2. Tokenization

[“How”, “do”, “I”, “install”, “a”, “new”, “package”, “with”, “pip”, “?”]

3. Stop Words & Lemmatization

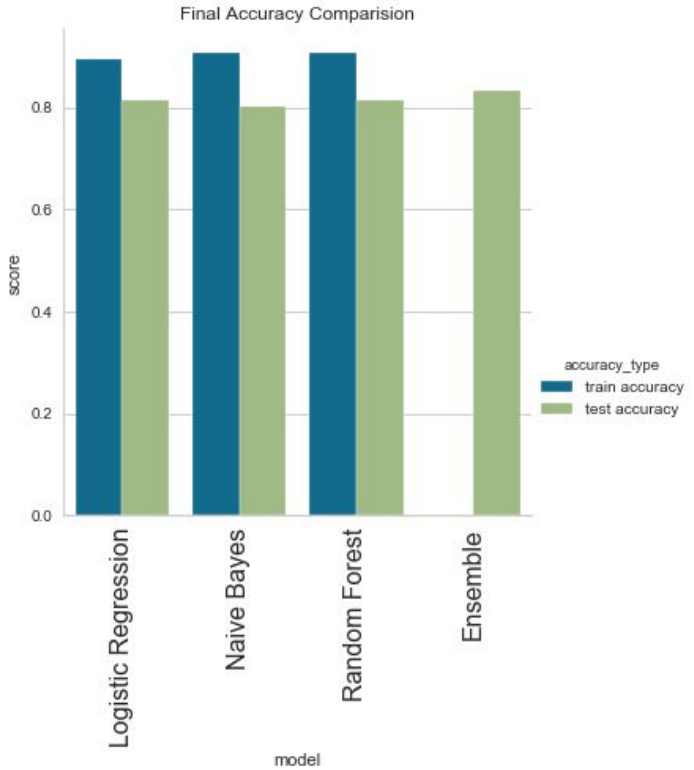
“install new package pip”

4. Vectorization (Sparse Matrix)

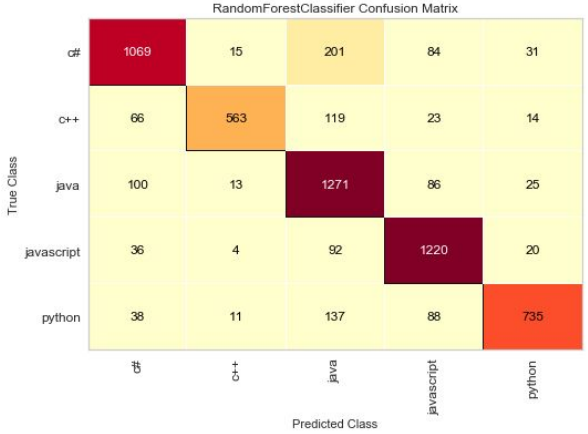
post id | word id | word weight across all documents

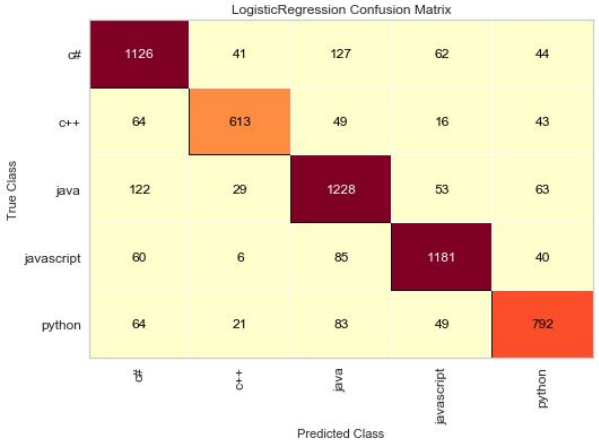
Model Selection

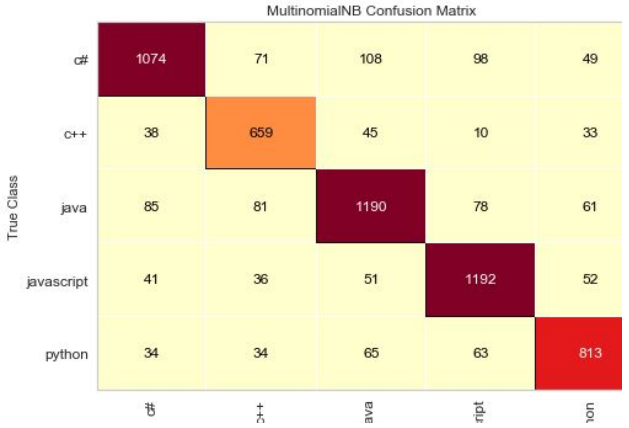
- Tested following classifiers:
  - Adaboost
  - Xgboost
  - Random Forest
  - Multinomial Naive Bayes
  - Logistic Regression
  - SVM classifier
- Random and Grid Search CV::
  - Performed on top 3 models shown in the bar chart
- Ensemble:
  - Majority vote among the 3 classifiers
- Metric of choice:
  - Accuracy



Confusion Matrix







	c# Pred	c++ Pred	java Pred	javascript Pred	python Pred
c# True	1123	39	129	74	35
c++ True	58	625	61	14	27
java True	88	27	1271	64	45
javascript True	37	12	75	1214	34
python True	43	13	87	60	806

Ensemble

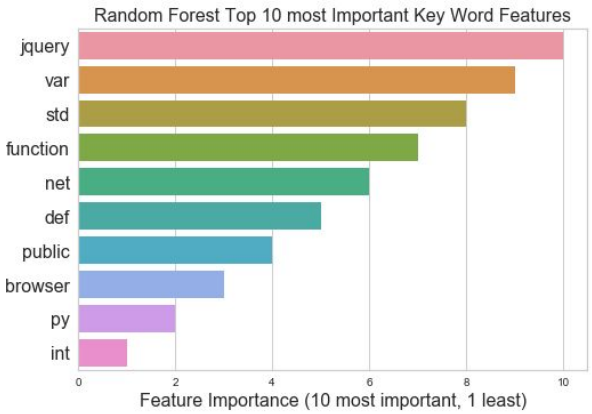
Top Words Per Category

	c#_lg	c++_lg	java_lg	javascript_lg	python_lg
0	writeline	boost	jvm	jquery	django
1	net	cout	jdk	backbone	numpy
2	dispose	std	jsp	console log	py
3	resharper	qt	system println	angularjs	def
4	window form	cpp	println	prototype	pythonic
5	linq	stl	spring	ecmascript	pep
6	msdn	gcc	jar	alert	matplotlib
7	ienumerable	int main	junit	browser	urllib
8	xaml	header file	hibernate	jslint	sqlalchemy
9	wpf	std string	jdbc	angular	typeerror

Logistic Regression

	c#_nb	c++_nb	java_nb	javascript_nb	python_nb
0	string	std	string	function	list
1	use	int	class	jquery	file
2	public	use	use	use	use
3	class	const	file	var	print
4	net	function	method	div	py
5	get	class	public	page	like
6	method	code	new	like	way
7	new	vector	get	html	self
8	code	foo	like	script	get
9	list	string	way	way	line

Naive Bayes



Random Forest

Clustering

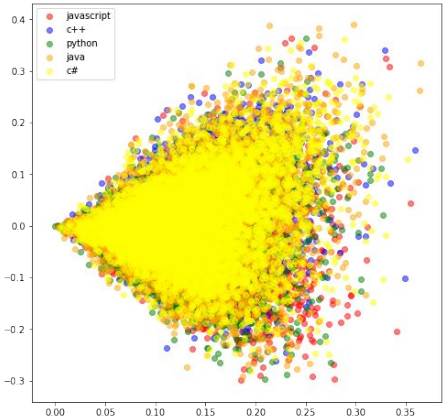
- Truncated SVD:
  - The 2 PCs explain only:
    - 0.00262252, 0.00623899 variabilities
  - Suggests lack of separation among texts
-
- KMeans Clustering:
  - silhouette_score: ~ 0
    - Indicate overlapping cluster s
- LDA:
  - Topics and body text from the topics suggest overlapping clusters
- 5 cluster may be the wrong number for hidden relationships

	topic_0	topic_1	topic_2	topic_3	topic_4
0	byte	android	std	org	log
1	long	script	foo	key	source
2	thread	item	event	foo	module
3	char	model	option	import	framework
4	size	date	template	self	org

BOW

	topic_0	topic_1	topic_2	topic_3	topic_4
0	script	date	std	foo	thread
1	byte	item	foo	key	log
2	event	script	vector	print	arraylist
3	long	android	template	log	json
4	image	button	date	bar	private

TFIDF



PC1 vs PC2

Try it For Yourself!

Stack Overflow Tag Predictor Web App

QUESTIONS?