STACKOVERFLOW SURVEY
NETWORK TOUR OF DATA SCIENCE (NTDS)

CLAUDIO LOUREIRO, CEM MUSLUOGLU, JORDAN WILLEMIN, CYRIL VAN SCHREVEN

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

TABLE OF CONTENTS

INTRODUCTION

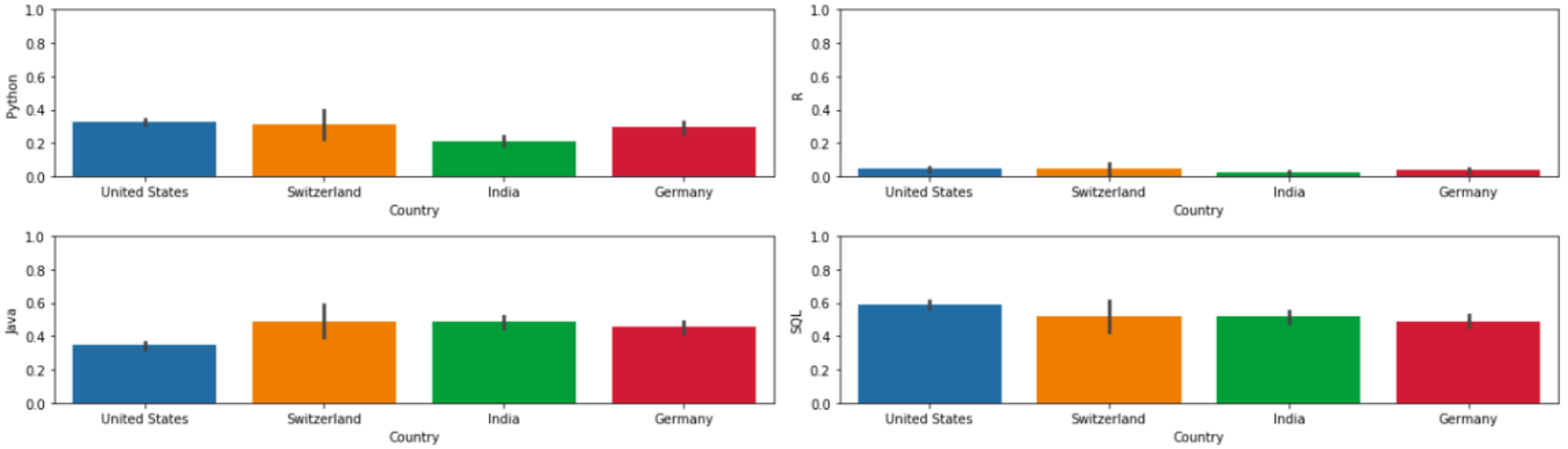
- Stackoverflow is a programming forum
- Users can fill in a form
- Questions of the survey
  - Different kinds (Country, formal education, career satisfaction, if people like debugging…)
  - From professional to personal
  - From concrete to abstract
- Does a correlation exist between features?

DATASET AND CLEANING

- Structure in a Panda dataframe
- Dataset
  - Feature = one asked question
  - Meaningful choice of features
- Data cleaning
  - Unanswered questions or «I prefer not to answer» are removed
  - Group data with equivalent answer
- Encode the programming language
  - Set to 1 the language if user knows to program it
  - Set to 0 otherwise

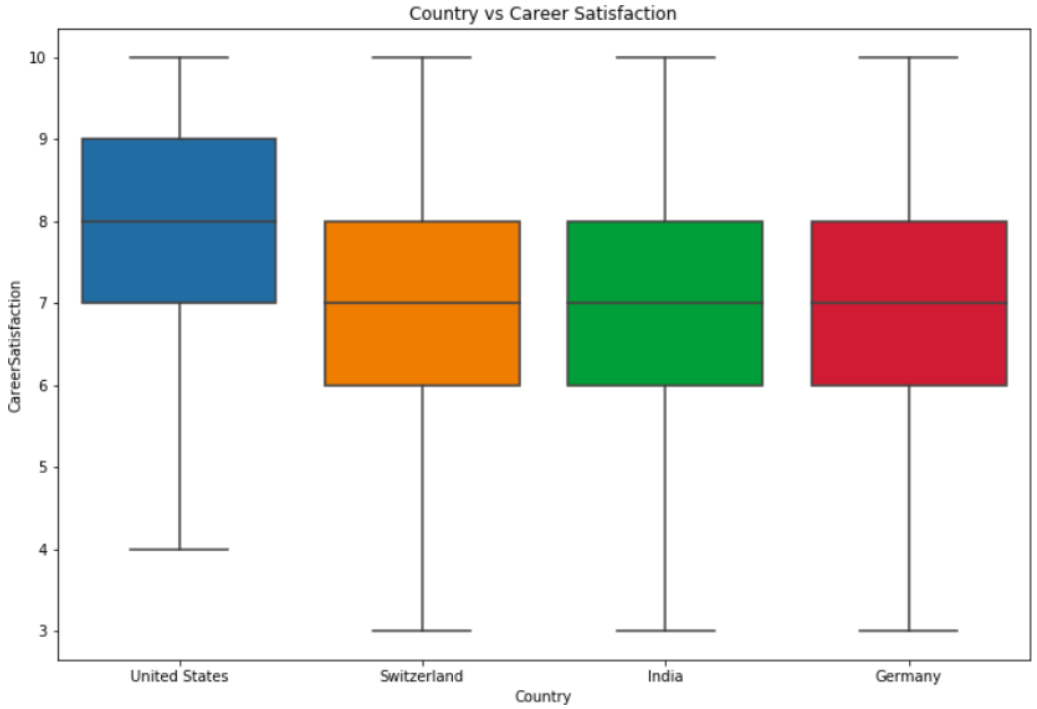
STATISTICAL STUDIES (1)

Country vs Programming language

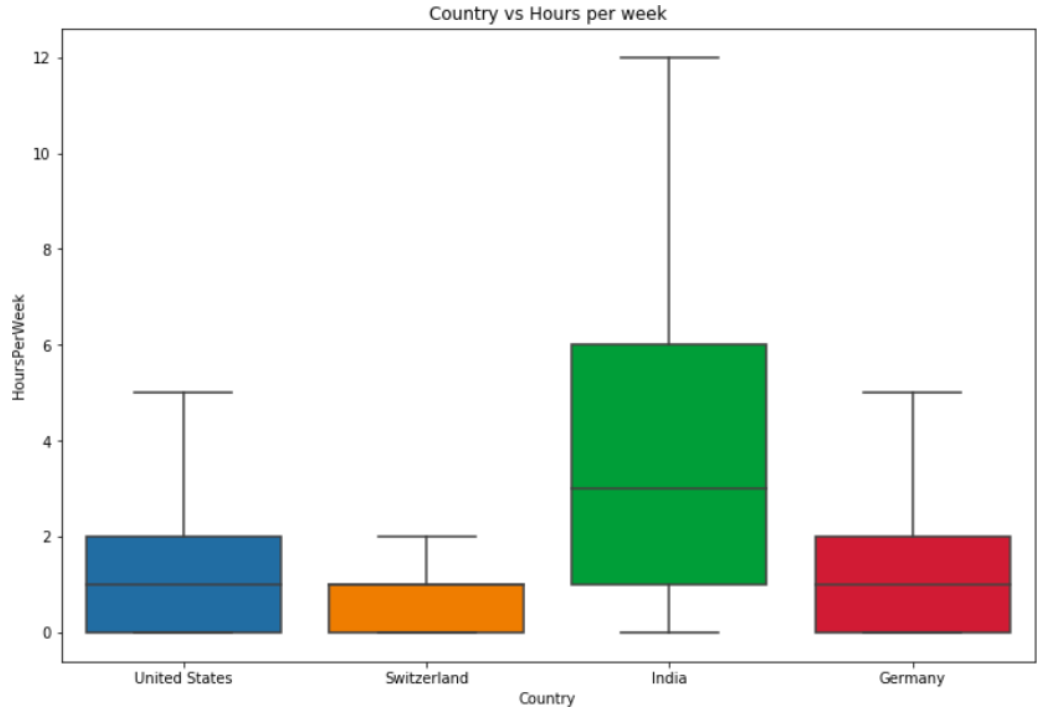


STATISTICAL STUDIES (2)

Country vs Career satisfaction



Country vs Hours per week

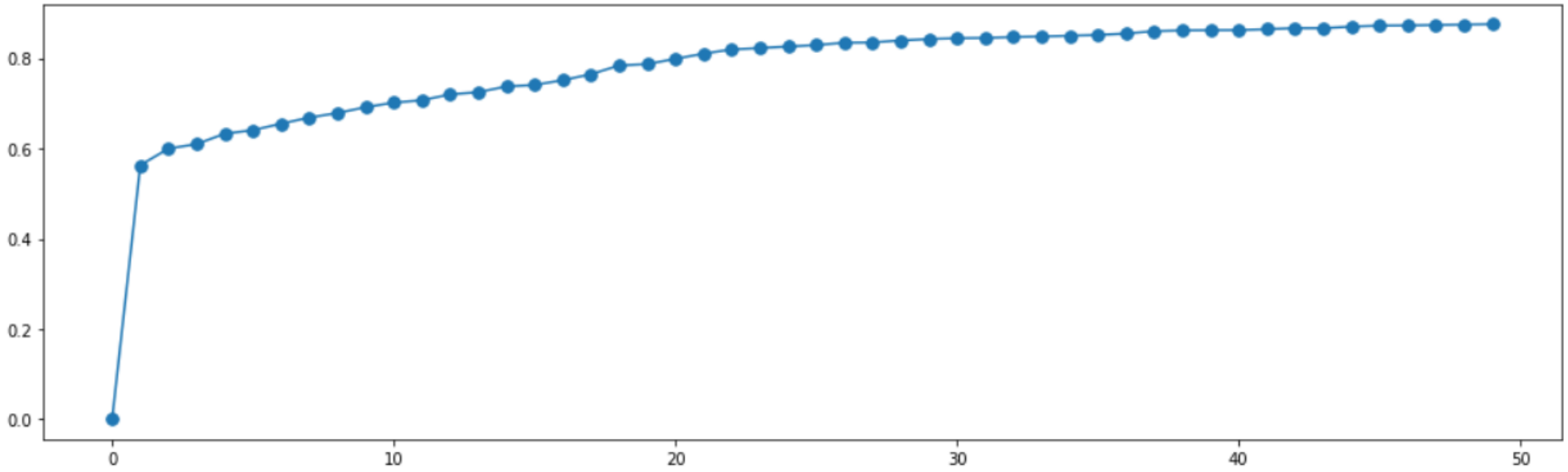


SPECTRAL GRAPH STUDY - PROCEDURE

- Check if a correlation exists between one tested feature and the chosen dataset
- The analyzed feature is removed from the dataset
- Calculation of the weight matrix
  - Nodes are the user
  - Create edge if users have the same answer to the same question
  - Weight of the edge is increased by one for each common answer
- Sparsification of the weight matrix
- Calculation of the Laplacian
- Visualization in 2D of the nodes representing the users

CASE 1: LOW CORRELATION (TESTED FEATURE: COUNTRY)

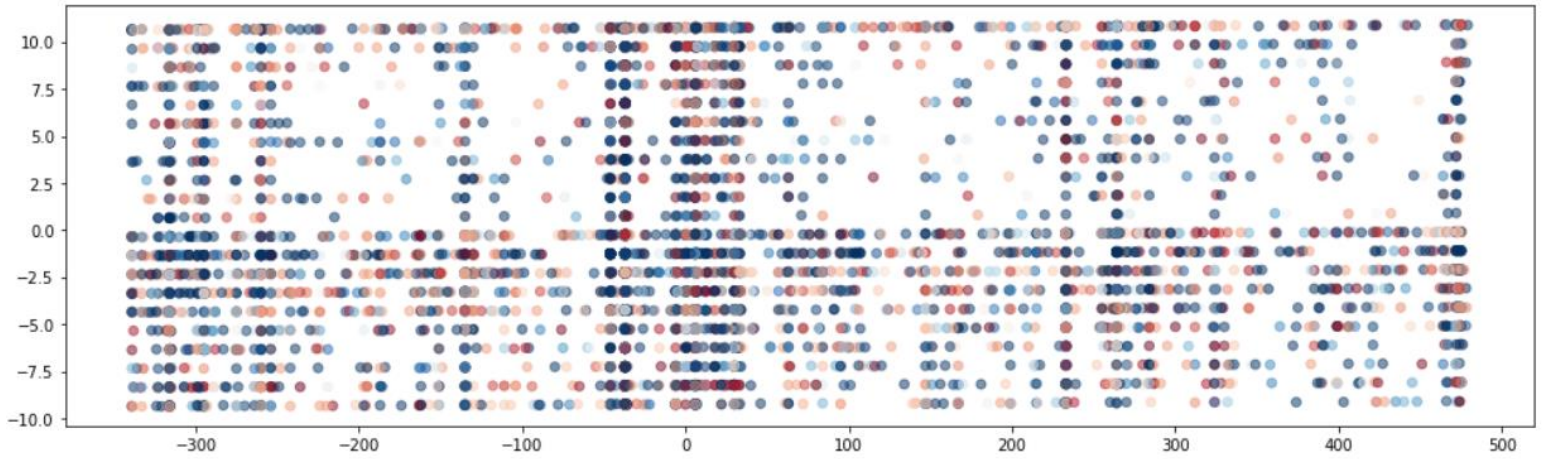
- Laplacian eigenvalues



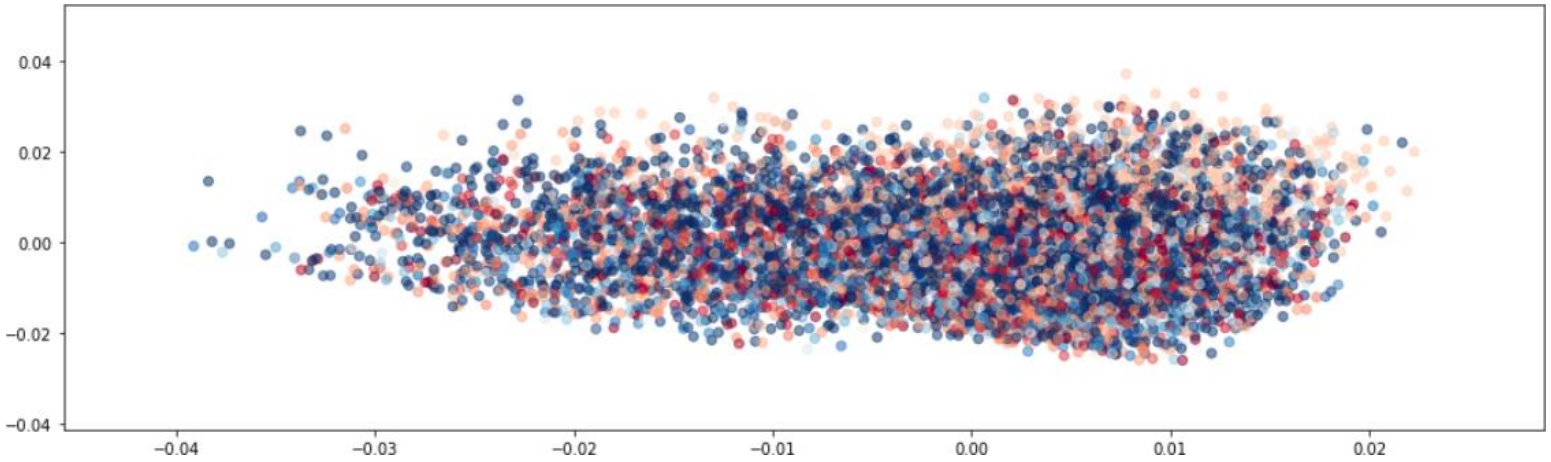
- Only one eigenvalue is null
- Thus we have a connected graph

CASE 1: LOW CORRELATION (TESTED FEATURE: COUNTRY)

- Principal component analysis (PCA)



- Graph Embedding

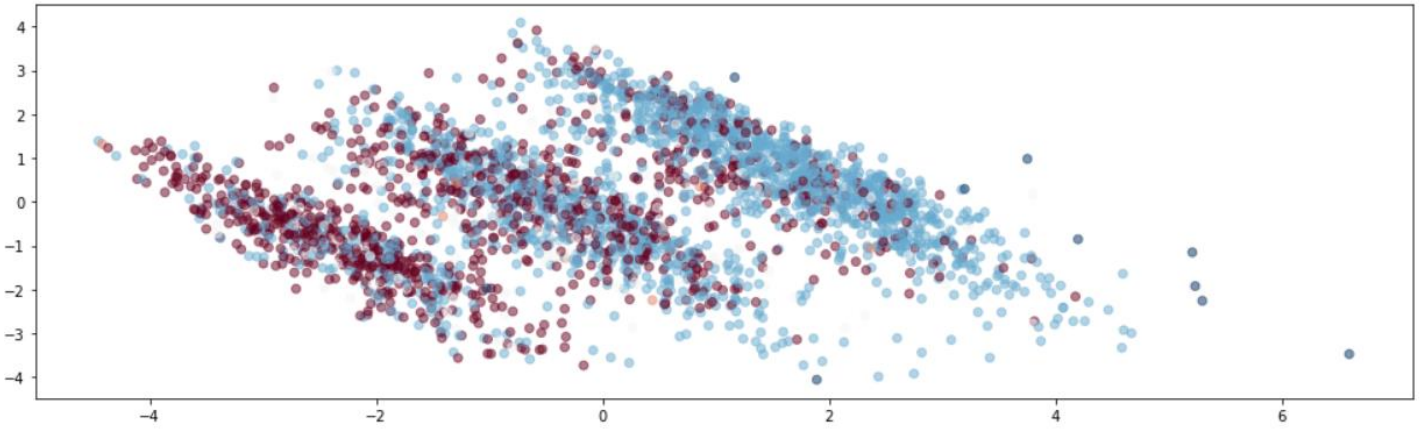


ANALYSIS

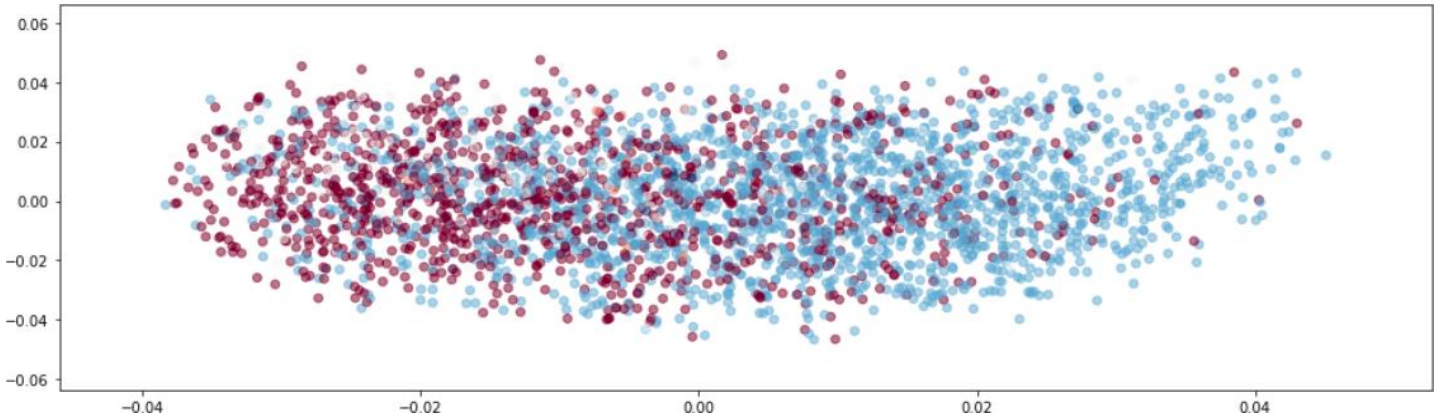
- The results are not satisfying
  - No clear clusters
- Many other tested features give such results i.e. “FormalEducation”, “CareerSatisfaction”, “EnjoyDebugging” …
- So we changed the initial set of features, s.t. they are less meaningful but potentially more correlated
- For example one chosen tested feature is “ProblemSolving” (Do you like solving problems ?)
- This latter is compared with the “ProgramHobby”, “BuildingThings”, “LearningNewTech”, “FriendsDevelopers”, “EnjoyDebugging” features

CASE 2: HIGH CORRELATION (TESTED FEATURE: PROBLEMSOLVING)

- Principal component analysis (PCA)



- Graph Embedding



MACHINE LEARNING

- Though we could not find obvious clusters for the "Country" feature we thought a correlation existed and tried to find it with a Machine Learning method
- We updated and increased the feature set again to test many at once
- The used data is cleaned again and separated in 3 groups: "Personal information", "Stackoverflow activity" and "Programming languages"
- We used a logistic regression with multinomial loss and validated by a cross validation
- The weight of each class has been balanced

MACHINE LEARNING - RESULTS

- Prediction accuracy :
  - Without features : $\frac{1}{Number\ of\ countries} = 7.7\%$
  - With "Personal information features" : 18.9%
  - With "Stackoverflow features" : 15.4%
  - With "Programming languages features" : 17.49%
  - With all features : 21%

MACHINE LEARNING – ANALYZING CORRELATIONS (1)

Features that helped the prediction

Country	1st important positive	2nd important positive	1st important negative	2nd important negative
Australia	BuildingThings	ChallengeMyself	LearningNewTech	SeriousWork
Brazil	SeriousWork	BuildingThings	KinshipDevelopers	AnnoyingUI
Canada	BuildingThings	AnnoyingUI	LearningNewTech	RightWrongWay
France	BuildingThings	DiversityImportant	ChallengeMyself	JobSecurity
Germany	SeriousWork	ProblemSolving	BuildingThings	ChallengeMyself
India	LearningNewTech	RightWrongWay	BuildingThings	ProblemSolving
Israel	SeriousWork	LearningNewTech	ProblemSolving	ChallengeMyself
Netherlands	KinshipDevelopers	SeriousWork	JobSecurity	BuildingThings
Poland	RightWrongWay	ChallengeMyself	ProblemSolving	DiversityImportant
Russian Federation	ChangeWorld	FriendsDevelopers	BuildingThings	DiversityImportant
Spain	BuildingThings	CompetePeers	ProblemSolving	AnnoyingUI
Sweden	ProblemSolving	InvestTimeTools	ChallengeMyself	JobSecurity
United Kingdom	BuildingThings	ChallengeMyself	SeriousWork	LearningNewTech

MACHINE LEARNING – ANALYZING CORRELATIONS (2)

Features that helped the prediction

Country	1st important positive	2nd important positive	1st important negative	2nd important negative
Australia	BuildingThings	C#	C	LearningNewTech
Brazil	SeriousWork	Java	Scala	KinshipDevelopers
Canada	BuildingThings	Python	LearningNewTech	StackOverflowAnswer
France	C	Scala	ChallengeMyself	C#
Germany	Java	R	BuildingThings	Scala
India	C	StackOverflowNewQuestion	C#	Scala
Israel	Scala	SeriousWork	PHP	StackOverflowJobListing
Netherlands	PHP	C#	C	Assembly
Poland	Scala	RightWrongWay	C	ProblemSolving
Russian Federation	Scala	Python	BuildingThings	Java
Spain	StackOverflowJobSearch	Java	Scala	C#
Sweden	ProblemSolving	StackOverflowCompanyPage	PHP	StackOverflowJobSearch
United Kingdom	C#	BuildingThings	Java	C

CONCLUSION

- First we tried to determine the countries where the users come from
- Unfortunately, there were not any obvious clusters but we did not give up !
- Thus we tried a Machine Learning approach
- We found that there was a slight correlation
- For predicting the countries, the features are probably not specific enough

QUESTIONS