Custom Content Delivery for Stack Overflow

Alok Kucheria

Raman Preet Singh

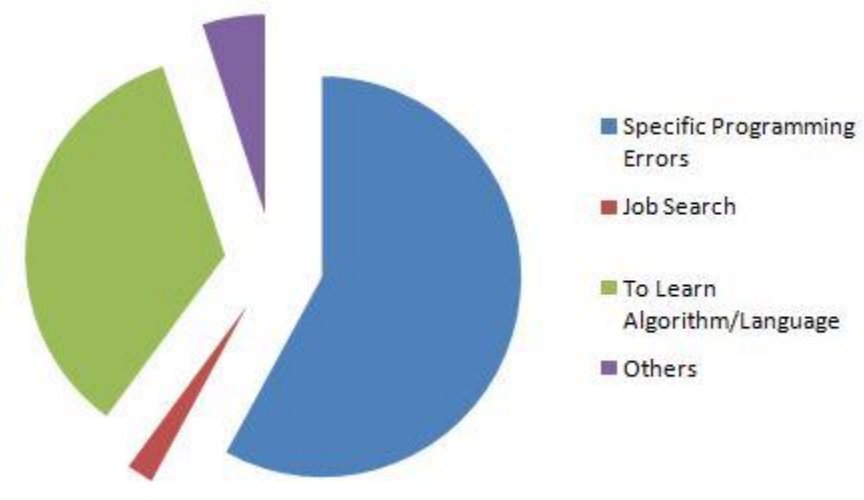
WHAT ?

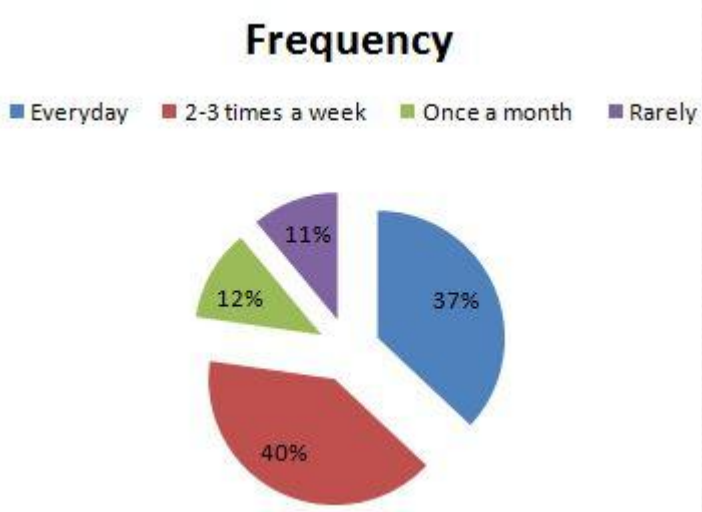
- A system to deliver customized content from Stack Overflow
- Customization is done as per user requirement
- Content is delivered in-mail as requested by the user

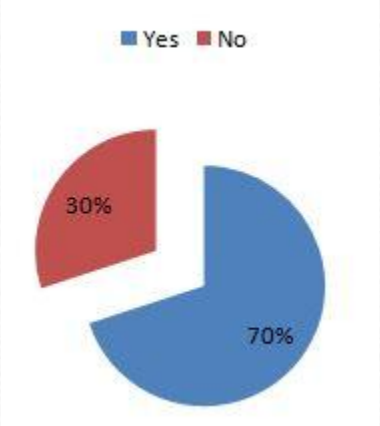
WHY ?

- Stack Overflow has a lot of information
- Currently, it is just present and available to users when they need it
- What if all this information could be utilized effectively by understanding user needs
- Information can be converted to knowledge

PROVE IT


Specific Programming Errors
Job Search
To Learn Algorithm/Language
Others


Frequency
Everyday
2-3 times a week
Once a month
Rarely
37%
40%
12%
11%


Yes
No
30%
70%

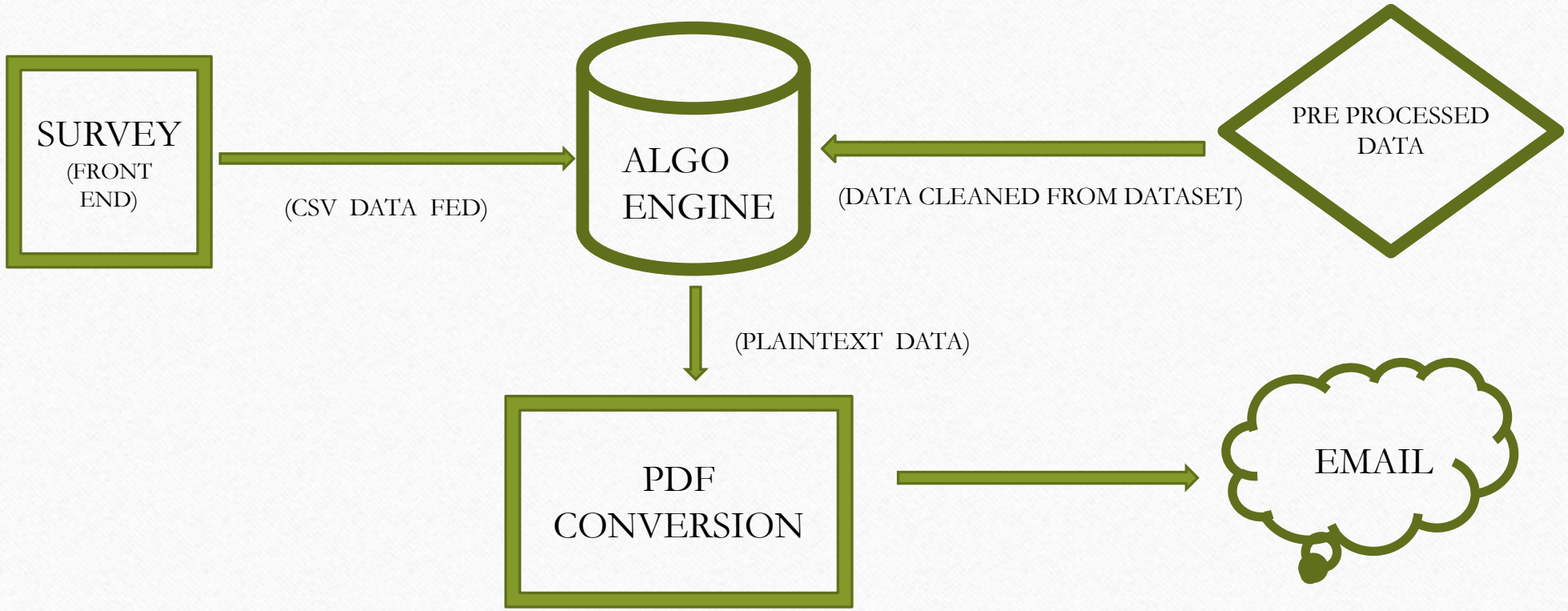
WHO WILL BENEFIT FROM THIS

- Any and everyone who uses StackOverflow
- All those people who wanted a newsletter from StackOverflow
- If you want to learn or just keep yourself with up to date.

HOW

- Simple Filtering
- Random Forest
- N-gram

THE MODEL


SURVEY
(FRONT END)
(CSV DATA FED)
ALGO ENGINE
(DATA CLEANED FROM DATASET)
PRE PROCESSED DATA
(PLAINTEXT DATA)
PDF CONVERSION
EMAIL

USING

- Archival data from https://archive.org/details/stackexchange
- Data provided by Stack Exchange for all its sites
- Used subset of programmers.stackexchange.com

Simple Filtering

- Based on user requirement, extract tags
- Based on tags, sort posts by frequency of votes, comments etc

The Good and Bad - Simple Filtering

- Advantages

Easy to implement, intuitive

- Disadvantages

Difficult to measure results, uncertainty based on user requirement

Random Forest

- Create multiple decision trees using content of tags and posts
- Split data into training and testing sets.
- Create feature vectors using tf-idf
- Build model and sort on probabilities

The Good and Bad - Random Forest

- Advantages

Accurate results, proven model based on probability

- Disadvantages

Depends on size of data set

N-grams

- Sequence of n-items from given text
- Used bigrams (n=2) for balance between performance and accuracy
- Uses content of posts rather than relying on tags

The Good and Bad - N-grams

- Advantages

Quality of results is high

- Disadvantages

High computational overhead

AND THE WINNER IS…

- N-grams!
- To the end-user, what matters is the quality of the content and not how difficult it was to extract it
- Since delivery can be weekly, bi-weekly etc, content can be generated using adequate resources

STATISTICS

- //CODE OWNERSHIP ISSUES

FUTURE WORKS

- Extend to entire data set
- Try web scraping instead of static data
- Implement as Stack Overflow app using Oauth
- Implementing a centralized database.

THANK YOU

Questions?