

Tag Recommendations for StackOverflow Posts


CSE 6240 Project Proposal

Ke Wang

Haochen Zhao

Tags and StackOverflow

- Tag: a type of meta-information
- Usage: indexing, categories, search terms, personalized bookmarks

 × 809538

a general-purpose programming language designed to be used in conjunction with the Java Virtual Machine (JVM). "Java

894 asked today, 5610 this week

 × 799890

JavaScript (not to be confused with Java or Unity3D's "JavaScript") is a dynamic weakly typed interpreted language typically

1023 asked today, 5753 this week

 × 763485

a multi-paradigm programming language encompassing strong typing, imperative, declarative, functional, generic, object-

688 asked today, 3792 this week

 × 714268

a popular general-purpose scripting language that is especially suited to web development. Fast, flexible and pragmatic,

744 asked today, 4470 this week

 × 638660

Google's OS for digital devices [Phone, Tablet, Auto, TV, Watch, Glass]. Please use Android-specific tags such as android-

721 asked today, 4421 this week

 × 586649

a popular cross-browser JavaScript library that facilitates DOM (Document Object Model - HTML Structure) traversal, event

539 asked today, 3111 this week

 × 396650


a dynamic and strongly typed programming language that is designed to emphasize usability. Two similar but

557 asked today, 3194 this week

 × 389620

the principal markup language used for structuring web pages and formatting content. The most recent revision to the

468 asked today, 2791 this week

 × 352037

a general-purpose programming language based on C. Use this tag for questions about code compiled with a C++ compiler,

356 asked today, 2104 this week

 × 313574


Apple's operating system for mobile devices, such as the iPhone, iPod touch, iPad and Apple TV (2nd generation and

407 asked today, 2416 this week

 × 306133

an open-source, relational database management system. If your issue relates to MySQLi, use the MySQLi tag instead.

296 asked today, 1879 this week

 × 287094













a style sheet language used for describing the look and formatting of HTML(Hyper Text Markup Language) and

352 asked today, 1966 this week

Problem

- Problem: When a user (maybe a new-comer) writes a new post, can we recommend some tags?

Title

B *I*            

Links Images Styling/Headers Lists Blockquotes Code HTML [advanced help »](#)

Tags

at least one tag such as (xml ruby r), max 5 tags

Traditional Approaches

- TF-IDF
 - Extract possible keywords from the post
 - Compute TF-IDF, and return top terms
- Similarity-based
 - Compute similarity based on TF-IDF
 - Recommend terms to the new post
- Tag-term co-occurrence
- LDA
$$P(c|d, \alpha, \beta) = \sum_{t=1}^T P(c|t, \beta)P(t|d, \alpha)$$

More Approaches

- pLSI-based and LDA-based Tag Recommender
 - Main idea: compare documents' topic distribution; assign tags to similar documents.

Datasets

- StackOverflow Dumps: posts, post links, tags, users
- ECML PKDD Discovery Challenge 2009

References

- [1] Zhang, N., Zhang, Y., & Tang, J. (2009). A tag recommendation system based on contents. *ECML PKDD Discovery Challenge 2009 (DC09)*, 285.
- [2] Xia, X., Lo, D., Wang, X., & Zhou, B. (2013, May). Tag recommendation in software information sites. *In Proceedings of the 10th Working Conference on Mining Software Repositories (pp. 287-296)*. IEEE Press.
- [3] Wang, S., Lo, D., Vasilescu, B., & Serebrenik, A. (2014, September). EnTagRec: An enhanced tag recommendation system for software information sites. *In Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on (pp. 291-300)*. IEEE.
- [4] Dredze, M., Wallach, H. M., Puller, D., & Pereira, F. (2008, January). Generating summary keywords for emails using topics. *In Proceedings of the 13th international conference on Intelligent user interfaces (pp. 199-206)*. ACM.