



CAEPIA 2018

Aplicaciones de la técnica de *topic model* en repositorios software

Carlos López Nozal, César García Osorio, Álar Arnaiz González y Mario Juez Gil

Universidad de Burgos - Área de Lenguajes y Sistemas Informáticos

Tabla de contenidos

1. Introducción
2. Definición de conceptos teóricos
3. Proceso de selección de artículos
4. Revisión bibliográfica
5. Conclusiones

Introducción

En la última década han surgido **forjas de proyectos software** de fácil acceso tanto para proyectos empresariales como para proyectos *OpenSource*

- SourceForge <https://sourceforge.net/>
- Github <https://github.com/>
- GitLab <https://about.gitlab.com/>
- Bitbucket <https://bitbucket.org/>


Estas forjas suelen integrar múltiples sistemas para dar soporte a los flujos de trabajo y registrar las interacciones textuales entre los miembros del equipo.

Introducción - Ejemplo revisión de código

[devtools] polish breakpoints doc #4136



Merged kaycebasques merged 9 commits into `google:master` from `unknown repository` on Feb 6, 2017

Conversation 28 Commits 9 Checks 0 Files changed 19 +266 -337



 kaycebasques commented on Feb 3, 2017 Member

- Updates screenshots to use the new standard.
- Uses consistent, general step-by-step instructions on how to set each type of breakpoint.
- Adds some overview information to top of doc explaining when to use each type.
- Updates link text and URLs in other docs.
- Changes the DOM change and exception breakpoint sections back to the standard format for guide sections. For a while I was experimenting with an interactive tutorial format for each of these sections. These are largely unnecessary now, now that the Get Started Debugging JS guide is shipped. It was also weird to have sections that were pretty much tutorials within a guide.
- Changes title of doc to "Breakpoints Guide". Eventually, all the DevTools docs will follow this format.
- Renames URL to `breakpoints` from `add-breakpoints` and redirects the old to the new. Kinda unnecessary, but I really like clean URLs

Reviewers

-  jpmmedley ✓
-  petele ●

Assignees

-  kaycebasques
-  jpmmedley

Labels

- P2
- cla: yes
- section-tools
- type-Content

Figure 1: Ejemplo real de una *pull request* disponible en <https://github.com/google/WebFundamentals/pull/4136>.

Una de las líneas de trabajo en el aprendizaje automático es el tratamiento de **grandes cantidades de documentos basados en texto** para poder extraer relaciones entre ellos.

El **modelado de temas (*topic modeling*)** es una técnica de **agrupamiento de documentos de texto (*clustering*)**.

Cada **tema se define siguiendo una distribución de probabilidades** de un conjunto de palabras que ocurren con más frecuencia en los documentos de ese tema

Introducción - Ejemplo de visualización de modelado de temas

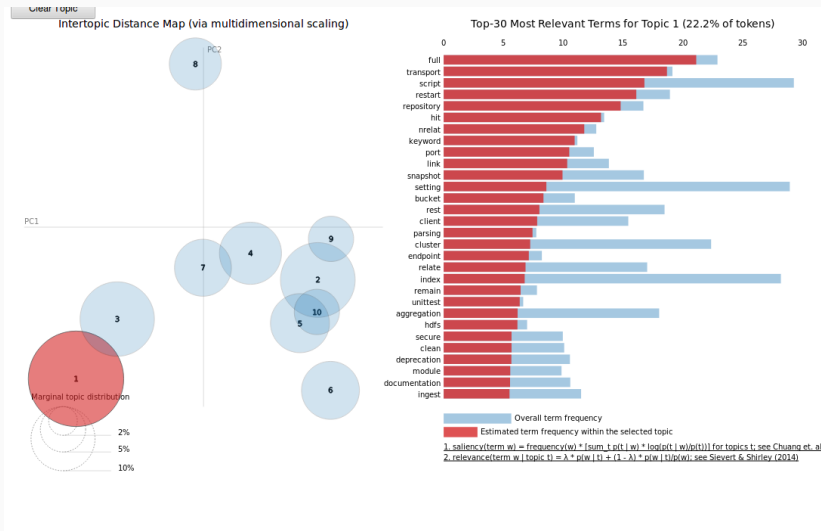


Figure 2: Ejemplo de visualización de resultados del modelado de temas. Basado en 4 303 pull request de ocho repositorios.

Objetivo

Estudiar en la literatura científica cómo se está aplicando la técnica de modelado de temas, en los conjuntos de datos extraídos desde repositorios software.

Definición de conceptos teóricos

Repositorios Software

Son espacios virtuales donde los equipos de desarrollo generan los artefactos colaborativos procedentes de las actividades de un proceso de desarrollo.

- **Sistemas de control de versiones** - commit, rama de desarrollo, centralizados o distribuidos
- **Sistemas de control de tareas** - issue, labels, pull request, identificadores
- **Sistemas de documentación** - Wiki, Readme, Web

Definición de conceptos teóricos - Agrupamiento y procesamiento de documentos

Modelado de temas o *topic modeling*

Es una técnica avanzada de recuperación de información que automáticamente encuentra los temas generales de en un conjunto de documentos de texto, llamado corpus, sin la necesidad de etiquetas, datos de entrenamiento o taxonomías predefinidas.

- **Algoritmos** - LSI (Latent Semantic Indexing), LDA (Latent Dirichlet Allocation) [2] y HDP (Hierarchical Dirichlet Process)
- **Palabras clave** - secuencia de palabras, documento, corpus, vocabulario, temas
- **Documento** d_i se puede modelar como una distribución multinomial θ^{d_i} sobre t temas, y cada **tema** z_k , $k = 1, \dots, t$ se modela como una distribución multinomial ϕ^k sobre el conjunto de palabras W
- **Bolsa de palabras (*Bag of Words*)** - Booleana, Raw TF, Escala logarítmica TF, IDF (*Inverse Document Frequency*, TF-IDF)

Definición de conceptos teóricos - Procesamiento del lenguaje natural

Procesamiento del lenguaje natural

Secuencia variable de **tareas de procesamiento** para la **extracción de palabras de los documentos**. Tareas: *tokenización*, lematización, selección de términos gramaticales (sustantivo, verbo, adjetivo, adverbio), eliminación de palabras sin significado e identificación de secuencias de palabras que se utilizan juntas.

Table 1: Ejemplo de tareas de procesamiento de lenguaje natural.

.token	.lemma	.tag	.es_alfabético	.palabra vacía
Updates	update	NOUN	True	False
link	link	VERB	True	False
text	text	NOUN	True	False
and	and	CONJ	True	True
URLs	url	NOUN	True	False
in	in	ADP	True	True
other	other	ADJ	True	True
docs	doc	NOUN	True	False

Proceso de selección de artículos

Proceso de selección de artículos

7	(TITLE-ABS-KEY ("Github" OR "Git" OR "GitLab" OR "Bitbucket") AND TITLE-ABS-KEY ("LDA" OR "LSI" OR "HDP" OR "Topic model"))	30 document results
6	(TITLE-ABS-KEY ("Github" OR "Git" OR "GitLab" OR "Bitbucket") AND TITLE-ABS-KEY ("LDA" OR "LSI" OR "HDP"))	22 document results
5	(TITLE-ABS-KEY (github) AND TITLE-ABS-KEY ("LDA" OR "LSI" OR "HDP"))	19 document results
4	(TITLE-ABS-KEY (github) AND TITLE-ABS-KEY ("LDA" OR "LSI"))	18 document results
3	(TITLE-ABS-KEY (github) AND TITLE-ABS-KEY (lda))	18 document results
2	TITLE-ABS-KEY (github)	3,697 document results

Figure 3: Resultados de la búsqueda sistemática en Scopus.

- Validación - lectura de los resúmenes y palabras clave para poder verificar la adecuación del artículo.
- Búsqueda no sistemática - *machine learning, text mining, text clustering, mining software repositories*

Table 2: Artículos seleccionados para la revisión

Título	Tipo	Año	Referencia
What are developers talking about? An analysis of topics and trends in Stack Overflow	Revista	2014	[1]
Open source is a continual bugfixing by a few	Actas congreso	2014	[3]
An insight into the pull requests of GitHub	Actas congreso	2014	[7]
Mining source code topics through topic model and words embedding	Actas congreso	2016	[12]
Topic-Based Integrator Matching for Pull Request	Actas congreso	2017	[6]
Cataloging GitHub repositories	Actas congreso	2017	[8]
Developer Identity Linkage and Behavior Mining Across GitHub and StackOverflow	Revista	2017	[10]
Mining developer behavior across git hub and stack overflow	Actas congreso	2017	[11]
Mining software repositories for defect categorization	Revista	2015	[5]
MSR4SM: Using topic models to effectively mining software repositories for software maintenance tasks	Revista	2015	[9]
Understanding Review Expertise of Developers: A Reviewer Recommendation Approach Based on Latent Dirichlet Allocation	Revista	2018	[4]

Revisión bibliográfica

Descripción de los conjuntos de datos

- **CD1 Tipo de entidades**, es una medida nominal que identifica los elementos de análisis dentro del repositorio software. Puede tomar los siguientes valores: **pull request, commit, source code, fichero Readme, issue, post**.
- **CD2 Número de entidades o documentos de texto** en la validación empírica del algoritmo de modelado de temas.
- **CD3 Número de repositorios** incluidos en el diseño experimental.
- **CD4 Uso de entidades de múltiples repositorios** es una medida booleana.
- **CD5 Disponibilidad** de acceso a los conjuntos de datos utilizados en la validación empírica es una medida nominal. *OpenSource* se ha utilizado para categorizar los trabajos con información de repositorios de tipo *OpenSource*.

Descripción de los conjuntos de datos

Table 3: Resumen de las características de los conjuntos de datos experimentales de la bibliografía.

BIB	CD1	CD2	CD3	CD4	CD5
[1]	post	3 447 987	7 meses actividad	SI	StackOverFlow
[3]	commits, issues	NO	43	NO	2014 MSR Challenge
[7]	pull request	9 421	78	NO	2014 MSR Challenge
[12]	source code	NO	100	SI	2013 MSR Challenge
[6]	pull request	4 364	3	NO	Github
[8]	fichero Readme	10 000	10 000	SI	Github
[11]	issue y post	16 000	No especificado	SI	Github StackOverFlow
[5]	Issue-bug	2 500	4	NO	OpenSource
[9]	post, commit, bugs	NO	3	NO	OpenSource
[4]	pull request	1 345	5	NO	Github

Aplicación del modelo de tópicos

- **AMT1 Algoritmo** es una medida nominal con el nombre del algoritmo de modelado de temas.
- **AMT2 Iteraciones** es un parámetro del algoritmo que sirve para determinar la condición de parada del algoritmo y convergencia de los resultados. Es un parámetro opcional.
- **AMT3 Número de tópicos/temas** es un parámetro necesario por el algoritmo LDA.
- **AMT4 Etiquetado manual** es una medida booleana.
- **AMT5 Validación** es una medida booleana que indica si en el diseño experimental se incluye alguna calibración de los parámetros del algoritmo.
- **AMT6 Transformación** del corpus es una medida nominal que indica la representación numérica del conjunto de documentos. Los posibles valores son: TF, IDF-TF.

Aplicación del modelo de tópicos

Table 4: Resumen de las características de la aplicación del algoritmo de modelado de la bibliografía.

BIB	AMT1	AMT2	AMT3	AMT4	AMT5	AMT6
[1]	LDA1	500	40	SI	NO	NO
[3]	LDA1	1000	50	NO	NO	NO
[7]	LDA2	3000	100	SI	NO	NO
[12]	LDA,EmbTE		NO	NO	SI	IDF -TF y TF
[6]	LDA2	1000	15	NO		NO
[8]	LDAGA	500	49	SI	SI	IDF-TF
[11]	LDA	NO	NO	NO	NO	NO
[5]	SDCL			NO		IDF-TF
[9]	LDAGA	NO	NO	NO	SI	NO
[4]	LDA1	NO	20	NO	NO	NO

Técnicas de procesamiento del lenguaje natural

- **TP1 Tokenización** es una medida ordinal que puede tomar los valores BAJO, MEDIO y ALTO.
- **TP2 Palabras vacías** es una medida booleana para indicar que se ha incluido esta tarea en el procesamiento del lenguaje.
- **TP3 Lematización** es una medida booleana para indicar que se ha incluido la tarea en el procesamiento del lenguaje.
- **TP4 Filtrado de palabras en el corpus** es una medida booleana para indicar que se ha incluido esta tarea en el procesamiento del lenguaje.
- **TP5 Etiquetado sintáctico** es una medida booleana para indicar que se ha incluido esta tarea en el procesamiento del lenguaje.
- **TP6 Identificación de secuencia de palabras** que se usan juntas es una medida booleana para indicar que se ha incluido esta tarea en el procesamiento del lenguaje.

Table 5: Resumen de las características de la aplicación del algoritmo de modelado de la bibliografía.

BIB	TP1	TP2	TP3	TP4	TP5	TP6
[1]	ALTO	SI	SI	NO	NO	2-grams
[3]	MEDIO	SI	NO	SI	NO	
[7]	BAJO	SI	SI	NO	NO	
[12]	MEDIO	SI	SI	SI	SI	
[6]	BAJO	SI	SI	NO	NO	
[8]	ALTO	SI	SI	NO	NO	
[11]	BAJO	SI1	SI	SI	NO	
[5]	BAJO	SI	SI	NO	NO	
[9]	MEDIO	SI	NO	NO	NO	
[4]	MEDIO	SI	SI	NO	NO	

Conclusiones

Conclusión general

Interés creciente en el uso de la técnica de minería de texto sobre documentos de los repositorios software. Su aplicación está orientada a mejorar la comprensión de las actividades de desarrollo.

- Caracterización conjunto de datos experimentales
 - Múltiples tipos de entidades de uno varios repositorios.
 - Número de instancias [1.345, 3.447.987].
 - 60% entidades del mismo repositorio.
 - Conjuntos de datos de proyectos Open Source, faltan datos de repositorios de empresas.

- Caracterización de la aplicación del algoritmo
 - LDA algoritmo de referencia.
 - 30% de los trabajos aplican una supervisión de la salida.
 - Configuración del algoritmo - iteraciones [500 – 3.000], número tópicos [15, 100].
 - La representación de BoW dominante es IDF-TF.
 - Solo el 30% se validan parámetros óptimos del algoritmo.
- Caracterización de técnicas de lenguaje natural
 - 100% de los trabajos aplican lematización y eliminación de palabras vacías.
 - 30% de los trabajos aplican filtrado de términos por frecuencia y análisis de elementos sintácticos.
 - 10% de los trabajos aplican identificación de secuencias de palabras.

Muchas gracias por su atención.



CAEPIA 2018

Aplicaciones de la técnica de *topic model* en repositorios software

Carlos López Nozal, César García Osorio, Álgar Arnaiz González y Mario Juez Gil

Universidad de Burgos - Área de Lenguajes y Sistemas Informáticos



A. Barua, S. Thomas, and A. Hassan.

What are developers talking about? an analysis of topics and trends in stack overflow.

Empirical Software Engineering, 19(3):619–654, 2014.



D. Blei, A. Ng, and M. Jordan.

Latent dirichlet allocation.

Journal of Machine Learning Research, 3(4-5):993–1022, 2003.



M. Fejzer, M. Wojtyna, M. Burzańska, P. Wiśniewski, and K. Stencel.

Open source is a continual bugfixing by a few.

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8716:153–162, 2014.



J. Kim and E. Lee.

Understanding review expertise of developers: A reviewer recommendation approach based on latent dirichlet allocation.

Symmetry, 10(4), 2018.



S. Kumaresh and R. Baskaran.

Mining software repositories for defect categorization.

Journal of Communications Software and Systems, 11(1):31–36, 2015.



Z. Liao, Y. Li, D. He, J. Wu, Y. Zhang, and X. Fan.

Topic-based integrator matching for pull request.

volume 2018-January, pages 1–6, 2018.



M. Rahman and C. Roy.

An insight into the pull requests of github.

pages 364–367, 2014.



A. Sharma, F. Thung, P. Kochhar, A. Sulistya, and D. Lo.

Cataloging github repositories.

volume Part F128635, pages 314–319, 2017.



X. Sun, B. Li, H. Leung, B. Li, and Y. Li.

Msr4sm: Using topic models to effectively mining software repositories for software maintenance tasks.

Information and Software Technology, 66:1–12, 2015.



Y. Xiong, Z. Meng, B. Shen, and W. Yin.

Developer identity linkage and behavior mining across github and stackoverflow.

International Journal of Software Engineering and Knowledge Engineering, 27(9-10):1409–1425, 2017.



Y. Xiong, Z. Meng, B. Shen, and W. Yin.

Mining developer behavior across git hub and stack overflow.

pages 578–583, 2017.



W. Zhang, Q. Sheng, E. Abebe, M. Ali Babar, and A. Zhou.

Mining source code topics through topic model and words embedding.

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10086 LNAI:664–676, 2016.