Docode 5: Building a scalable plagiarism detection system

Gaspar Pizarro V.

Web Intelligence Centre

November 22, 2016

Outline

The core algorithm

- Suspect vs source
- Based on word n-grams (shingles)

```
this module works as the user
interface
```

The core algorithm

- Suspect vs source
- Based on word n-grams (shingles)

~~this~~ module works ~~as~~ ~~the~~ user interface

The core algorithm

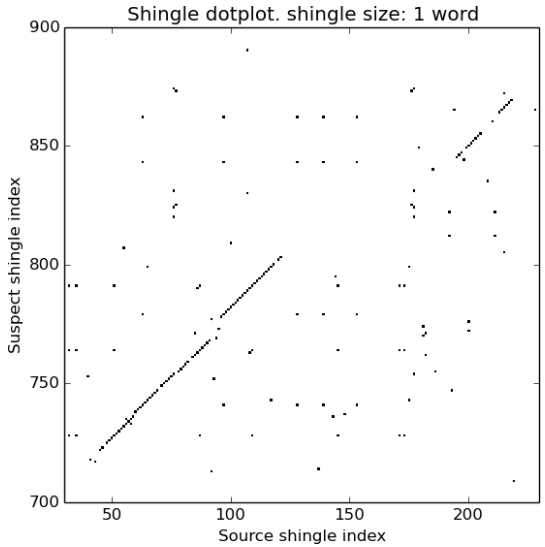
- Suspect vs source
- Based on word n-grams (shingles)

`module works user interface`

- `[(module), (works), (user), (interface)]`
- `[(module works), (works user), (user interface)]`
- `[(module works user), (works user interface)]`

The core algorithm

- Suspect vs source
- Based on word n-grams (shingles)

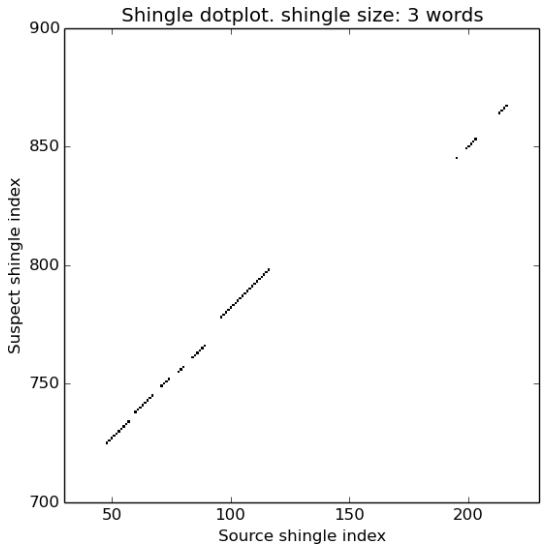


`module works user interface`

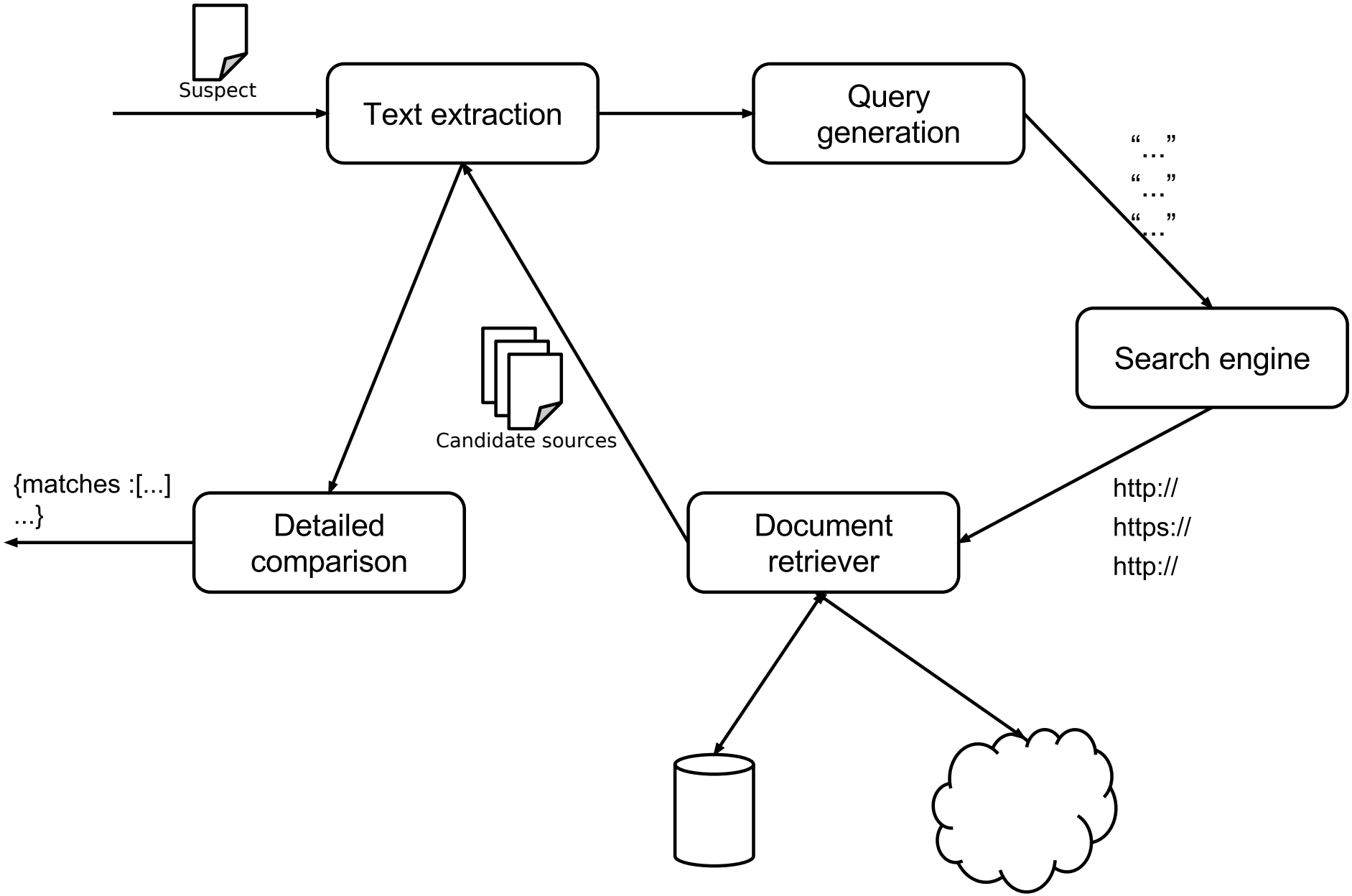
- `[(module), (works), (user), (interface)]`
- `[(module works), (works user), (user interface)]`
- `[(module works user), (works user interface)]`

Dotplot

- Find lines
- Merge closest ones
- Delete smallest ones



Web/Repository analysis



Web/Repository analysis: Example

`the_suspect.pdf`

Web/Repository analysis: Example

Chile colonial fue un periodo de la historia ubicada entre Fines del Siglo XVI y Principios del Siglo XIV, Esta época se caracterizó por la creación de una organización institucional muy compleja, el mestizaje y la combinación cultural que darían origen a la sociedad chilena como tal. La colonia Tuvo varios aspectos y características, como la forma de administrar Chile; Los sistemas económicos que eran antiguamente utilizados en España también serían empleados en América, ver qué sería lo que traería más productividad y provecho a nuestra colonia; el sistema educacional que sería empleado, en algunos casos los de mayor situación económica tenían más oportunidades y opciones de estudio que los demás, y como fue la estructura social de esos tiempos, Quienes salían favorecidos por el rango social en el que se encontraban, y quienes eran discriminados y afectados por el bajo rango social; esto produciría desacuerdos en opiniones y complicaciones entre la población que en ese entonces habitaba América, también hubo un factor muy importante que influyo en todos estos aspectos, que fue la religión tenia influencia en todo tipo de decisión En este periodo también se formó una gran combinación racial y cultural, que se originó con la convivencia de españoles e indígenas.

Web/Repository analysis: Example

```
Chile periodo historia Fines Siglo Principios Siglo XIV época creación

organización mestizaje combinación origen sociedad colonia aspectos características forma
Chile

sistemas España empleados América productividad provecho colonia sistema casos situación

oportunidades opciones estudio estructura tiempos favorecidos rango rango desacuerdos opiniones

complicaciones población América factor aspectos religión influencia tipo decisión periodo

combinación convivencia españoles indígenas
```

Web/Repository analysis: Example

```
https://es.wikipedia.org/wiki/Historia_de_Chile

https://es.wikipedia.org/wiki/Mestizaje_en_Am%C3%A9rica

https://es.wikipedia.org/wiki/Econom%C3%ADa_de_Espa%C3%B1a

https://es.scribd.com/doc/57536530/Estructura-de-Planes-y-Programas.pdf

https://es.wikipedia.org/wiki/Gripe

http://www.pps.k12.or.us/district/depts/edmedia/videoteca/curso3/htmlb/SEC_51.HTM
```

Web/Repository analysis: Example

https://es.wikipedia.org/wiki/Historia_de_Chile

~~https://es.wikipedia.org/wiki/Mestizaje_en_Am%C3%A9rica~~

https://es.wikipedia.org/wiki/Econom%C3%ADa_de_Espa%C3%B1a

https://es.scribd.com/doc/57536530/Estructura-de-Planes-y-Programas.pdf

~~https://es.wikipedia.org/wiki/Gripe~~

~~http://www.pps.k12.or.us/district/depts/edmedia/videoteca/curso3/htmlb/SEC_51.HTM~~

Web/Repository analysis: Example

Web/Repository analysis: Example

El periodo prehispánico corresponde a la historia de las diferentes etnias amerindias presentes en el territorio, extendiéndose desde alrededor del año 14 800 a. C. hasta la llegada de los españoles. A partir de 1492, se iniciaron las exploraciones europeas en el continente americano. En 1520 Fernando de Magallanes y su expedición fueron los primeros europeos en llegar a Chile por el sur a través del estrecho que hoy lleva su nombre, y en 1536 Diego de Almagro comandó una expedición hasta el Valle del Aconcagua y el norte del actual Chile.

La economía de España es la quinta por tamaño en la Unión Europea y la decimotercera a nivel mundial en términos nominales.17 En términos relativos o de paridad de poder adquisitivo, se encuentra también entre las mayores del mundo (ver Anexo:Países por PIB (PPA)). Según un informe de The Economist, España es el 10º país del mundo con mayor calidad de vida.18 Como en la economía de todos los países europeos, el sector terciario o sector servicios es el que tiene un mayor peso. La moneda de España es, desde 2002, el euro.

La principal estrategia para la consecución de este objetivo en educación básica plantea realizar una reforma integral de la educación básica, centrada en la adopción de un modelo educativo basado en competencias que responda a las necesidades de desarrollo de México en el siglo XXI con miras a lograr mayor articulación y eficiencia entre preescolar, primaria y secundaria

Web/Repository analysis: Example

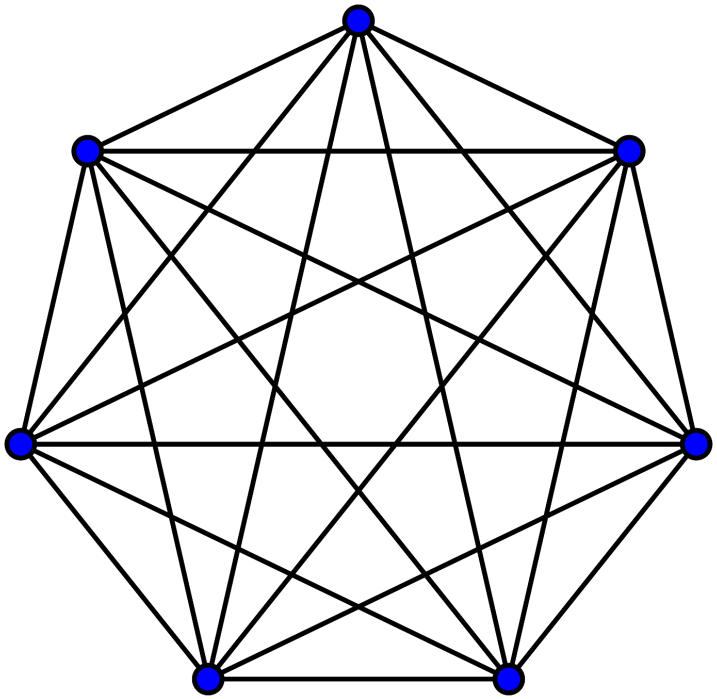
El periodo prehispánico corresponde a la historia de las diferentes etnias amerindias presentes en el territorio, extendiéndose desde alrededor del año 14 800 a. C. hasta la llegada de los españoles. A partir de 1492, se iniciaron las exploraciones europeas en el continente americano. En 1520 Fernando de Magallanes y su expedición fueron los primeros europeos en llegar a Chile por el sur a través del estrecho que hoy lleva su nombre, y en 1536 Diego de Almagro comandó una expedición hasta el Valle del Aconcagua y el norte del actual Chile.

La economía de España es la quinta por tamaño en la Unión Europea y la decimotercera a nivel mundial en términos nominales.17 En términos relativos o de paridad de poder adquisitivo, se encuentra también entre las mayores del mundo (ver Anexo:Países por PIB (PPA)). Según un informe de The Economist, España es el 10º país del mundo con mayor calidad de vida.18 Como en la economía de todos los países europeos, el sector terciario o sector servicios es el que tiene un mayor peso. La moneda de España es, desde 2002, el euro.

La principal estrategia para la consecución de este objetivo en educación básica plantea realizar una reforma integral de la educación básica, centrada en la adopción de un modelo educativo basado en competencias que responda a las necesidades de desarrollo de México en el siglo XXI con miras a lograr mayor articulación y eficiencia entre preescolar, primaria y secundaria

Group analysis

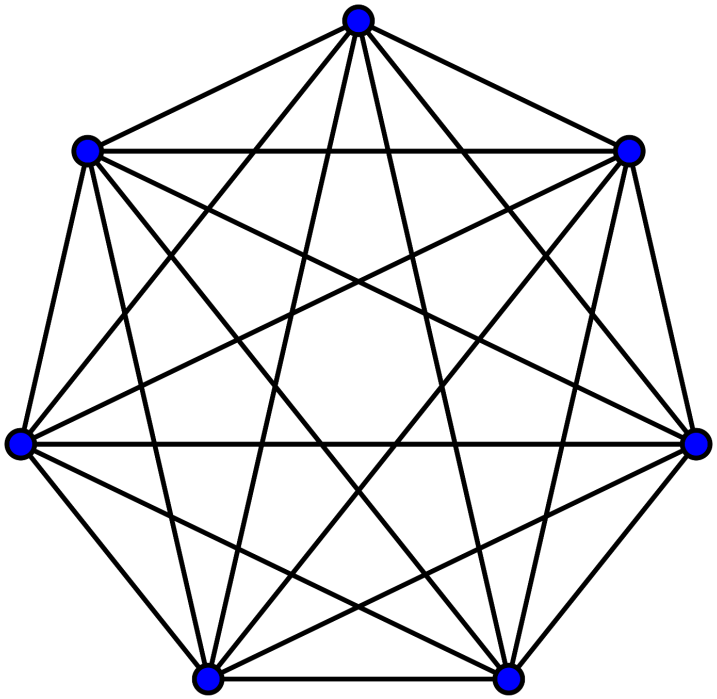
Brute force



- Grows like n^2
- Preprocessing

Group analysis

Brute force



- Grows like n^2
- Preprocessing
- It can be improved

Group analysis

Inverted index

Doc1

El periodo prehispánico corresponde a la historia de las diferentes etnias amerindias presentes en el territorio, extendiéndose desde alrededor del año 14 800 a. C. hasta la llegada de los españoles. A partir de 1492, se iniciaron las exploraciones europeas en el continente americano. En 1520 Fernando de Magallanes y su expedición fueron los primeros europeos en llegar a Chile por el sur a través del estrecho que hoy lleva su nombre, y en 1536 Diego de Almagro comandó una expedición hasta el Valle del Aconcagua y el norte del actual Chile.

Doc2

La economía de España es la quinta por tamaño en la Unión Europea y la decimotercera a nivel mundial en términos nominales.17 En términos relativos o de paridad de poder adquisitivo, se encuentra también entre las mayores del mundo (ver Anexo:Países por PIB (PPA)). Según un informe de The Economist, España es el 10° país del mundo con mayor calidad de vida.18 Como en la economía de todos los países europeos, el sector terciario o sector servicios es el que tiene un mayor peso. La moneda de España es, desde 2002, el euro.

Doc3

La principal estrategia para la consecución de este objetivo en educación básica plantea realizar una reforma integral de la educación básica, se encuentra centrada en la adopción de un modelo educativo basado en competencias que responda a las necesidades de desarrollo de México en el siglo XXI con miras a lograr mayor articulación y eficiencia entre preescolar, primaria y secundaria

Group analysis

Inverted index

Words to documents

periodo	→	Doc1, 3
España	→	Doc2, 15, 301, 495
Encuentra	→	Doc2, 186 | Doc3, 146

Group analysis

Inverted index

Words to documents

periodo	→	Doc1, 3
España	→	Doc2, 15, 301, 495
Encuentra	→	Doc2, 186 | Doc3, 146

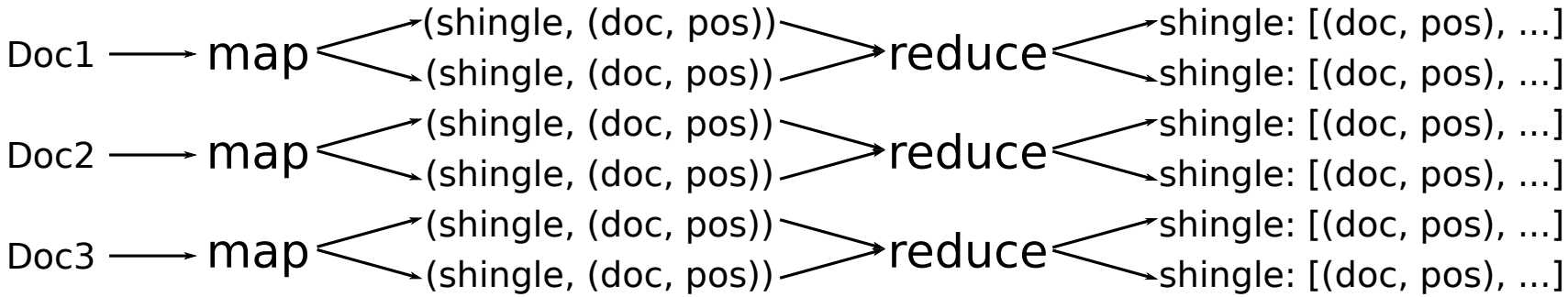
Differences

- Shingles instead of words
- Position in list of shingles instead of position in string

Group analysis

MapReduce, two passes

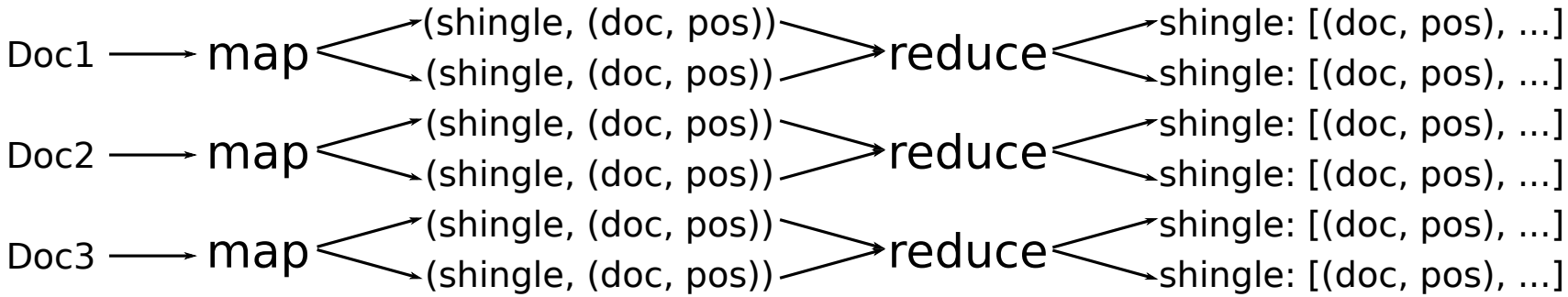
- Indexing job



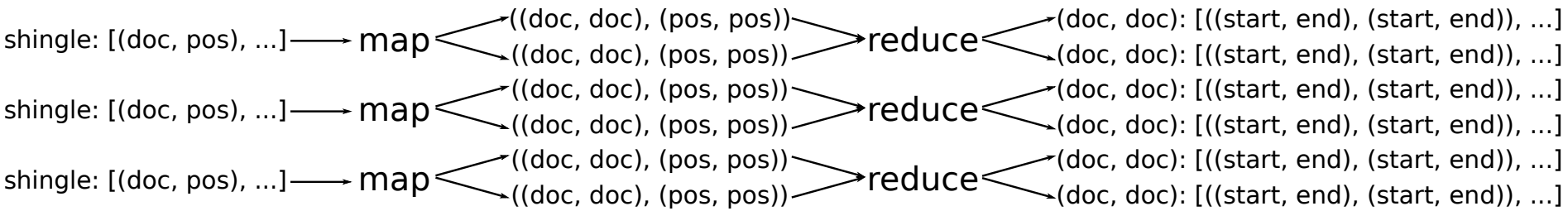
Group analysis

MapReduce, two passes

- Indexing job

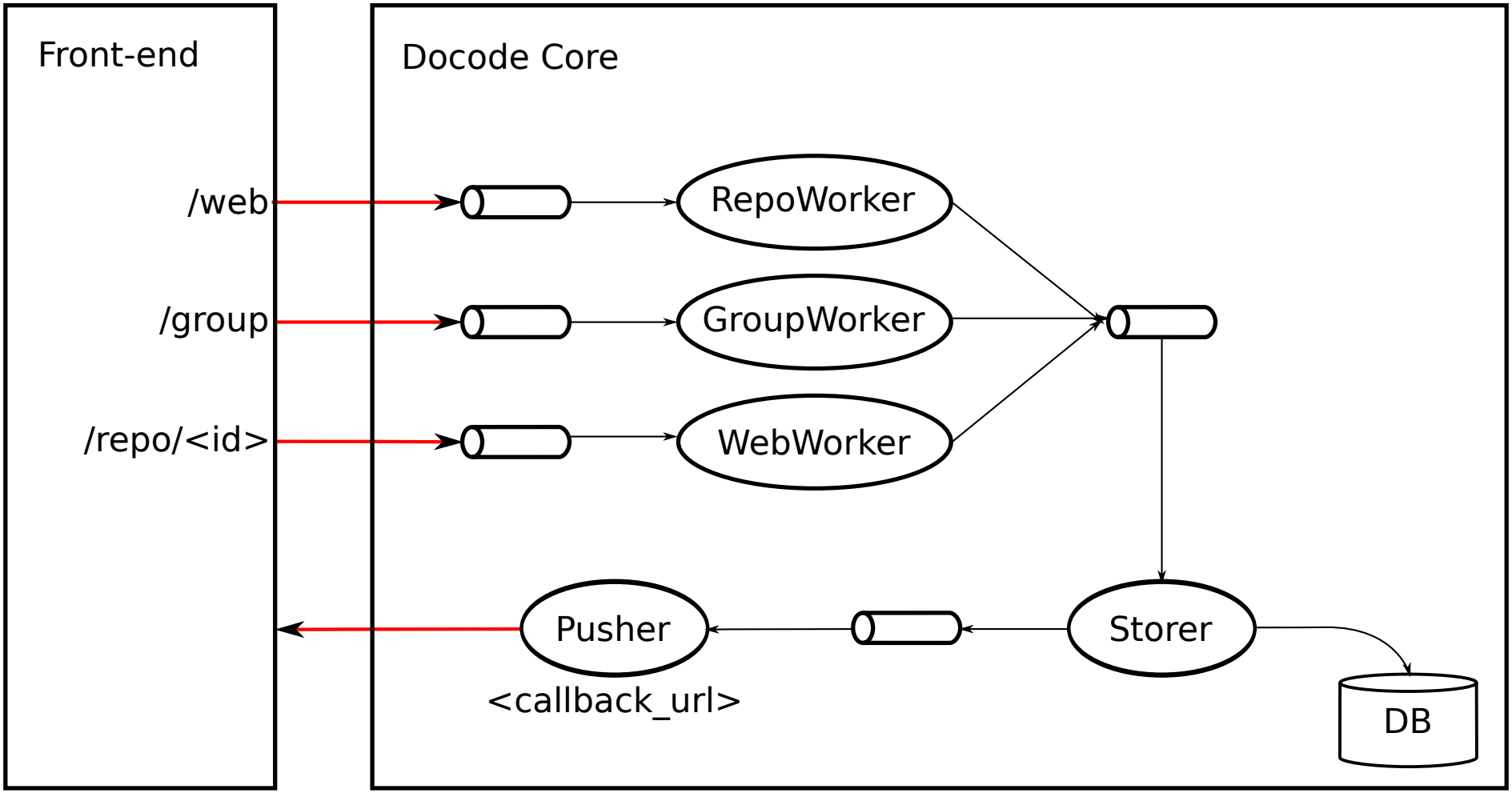


- Matching job

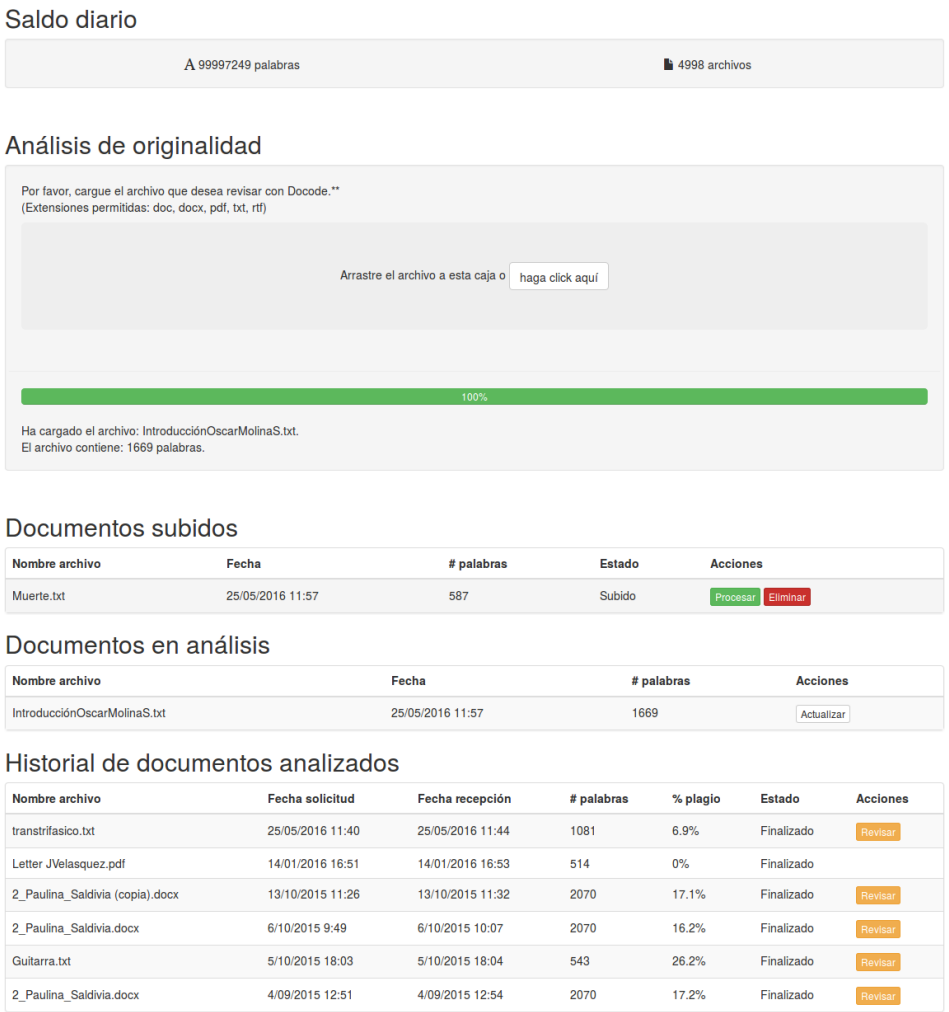


Outline

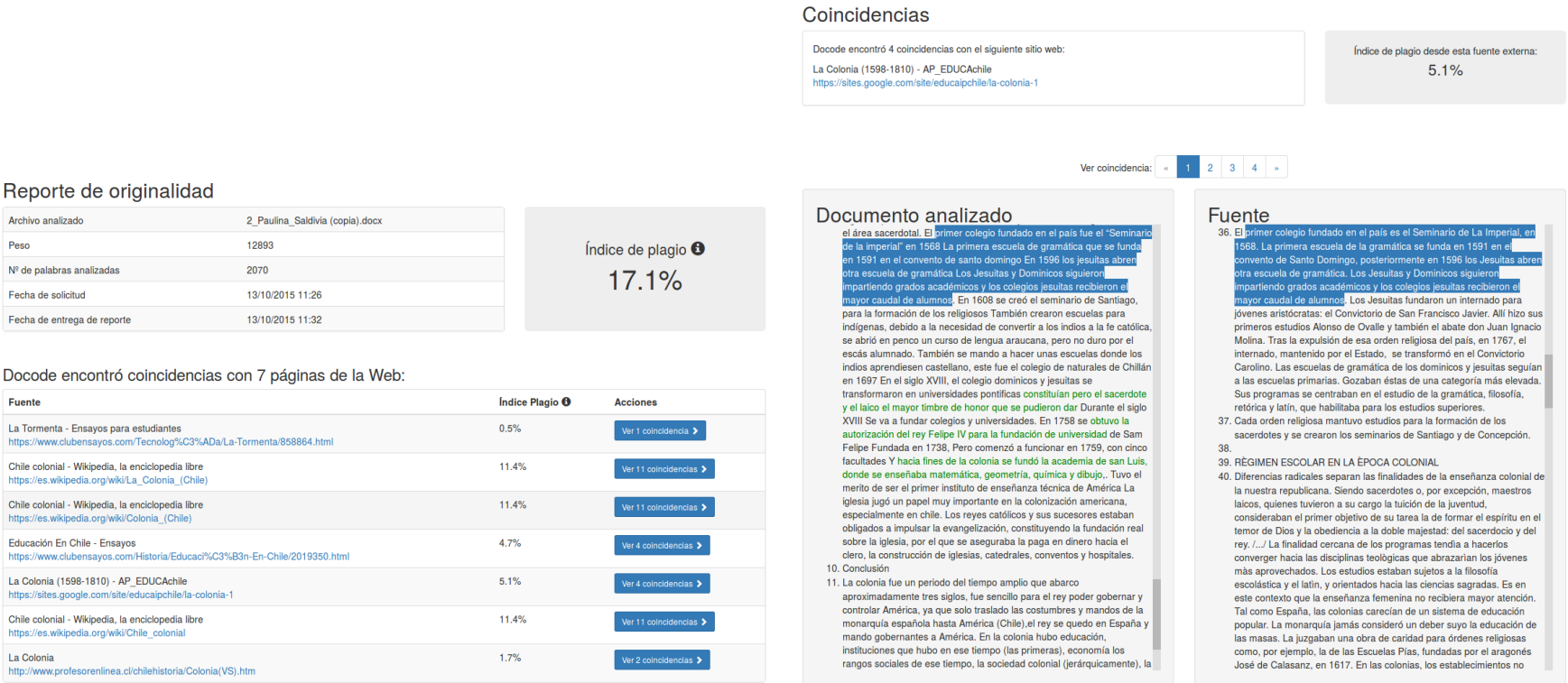
High-level architecture



Front-end lite



Front-end lite



Coincidencias

Docode encontró 4 coincidencias con el siguiente sitio web:

La Colonia (1598-1810) - AP_EDUCAchile
https://sites.google.com/site/educaipchile/la-colonia-1

Índice de plagio desde esta fuente externa:
5.1%

Ver coincidencia: « 1 2 3 4 »

Reporte de originalidad

Archivo analizado	2_Paulina_Saldivia (copia).docx
Peso	12893
Nº de palabras analizadas	2070
Fecha de solicitud	13/10/2015 11:26
Fecha de entrega de reporte	13/10/2015 11:32

Índice de plagio
17.1%

Docode encontró coincidencias con 7 páginas de la Web:

Fuente	Índice Plagio	Acciones
La Tormenta - Ensayos para estudiantes https://www.clubensayos.com/Tecnolog%C3%ADa/La-Tormenta/858864.html	0.5%	Ver 1 coincidencia
Chile colonial - Wikipedia, la enciclopedia libre https://es.wikipedia.org/wiki/La_Colonia_(Chile)	11.4%	Ver 11 coincidencias
Chile colonial - Wikipedia, la enciclopedia libre https://es.wikipedia.org/wiki/Colonia_(Chile)	11.4%	Ver 11 coincidencias
Educación En Chile - Ensayos https://www.clubensayos.com/Historia/Educaci%C3%B3n-En-Chile/2019350.html	4.7%	Ver 4 coincidencias
La Colonia (1598-1810) - AP_EDUCAchile https://sites.google.com/site/educaipchile/la-colonia-1	5.1%	Ver 4 coincidencias
Chile colonial - Wikipedia, la enciclopedia libre https://es.wikipedia.org/wiki/Chile_colonial	11.4%	Ver 11 coincidencias
La Colonia http://www.profesorenlinea.cl/chilehistoria/Colonia(VS).htm	1.7%	Ver 2 coincidencias

Documento analizado

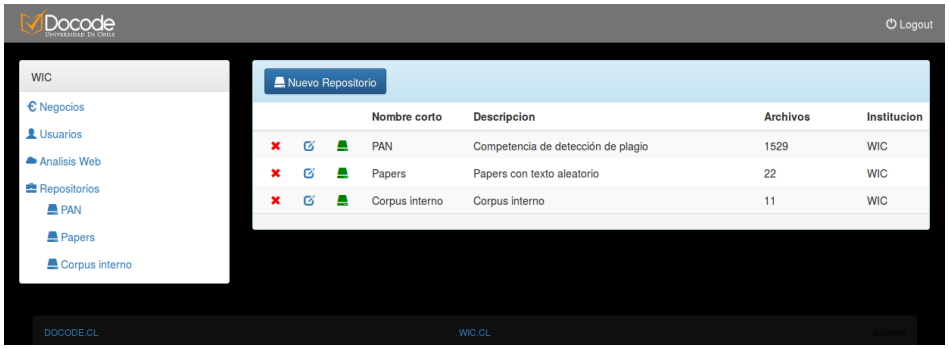
el área sacerdotal. El primer colegio fundado en el país fue el "Seminario de la imperial" en 1568 La primera escuela de gramática que se funda en 1591 en el convento de santo domingo En 1596 los jesuitas abren otra escuela de gramática Los Jesuitas y Dominicos siguieron impartiendo grados académicos y los colegios jesuitas recibieron el mayor caudal de alumnos. En 1608 se creó el seminario de Santiago, para la formación de los religiosos También crearon escuelas para indígenas, debido a la necesidad de convertir a los indios a la fe católica, se abrió en penco un curso de lengua araucana, pero no duro por el escás alumnado. También se mando a hacer unas escuelas donde los indios aprendiesen castellano, este fue el colegio de naturales de Chillán en 1697 En el siglo XVIII, el colegio dominicos y jesuitas se transformaron en universidades pontificas constituían pero el sacerdote y el laico el mayor timbre de honor que se pudieron dar Durante el siglo XVIII Se va a fundar colegios y universidades. En 1758 se obtuvo la autorización del rey Felipe IV para la fundación de universidad de Sam Felipe Fundada en 1738, Pero comenzó a funcionar en 1759, con cinco facultades Y hacia fines de la colonia se fundó la academia de san Luis, donde se enseñaba matemática, geometría, química y dibujo,. Tuvo el merito de ser el primer instituto de enseñanza técnica de América La iglesia jugó un papel muy importante en la colonización americana, especialmente en chile. Los reyes católicos y sus sucesores estaban obligados a impulsar la evangelización, constituyendo la fundación real sobre la iglesia, por el que se aseguraba la paga en dinero hacia el clero, la construcción de iglesias, catedrales, conventos y hospitales.

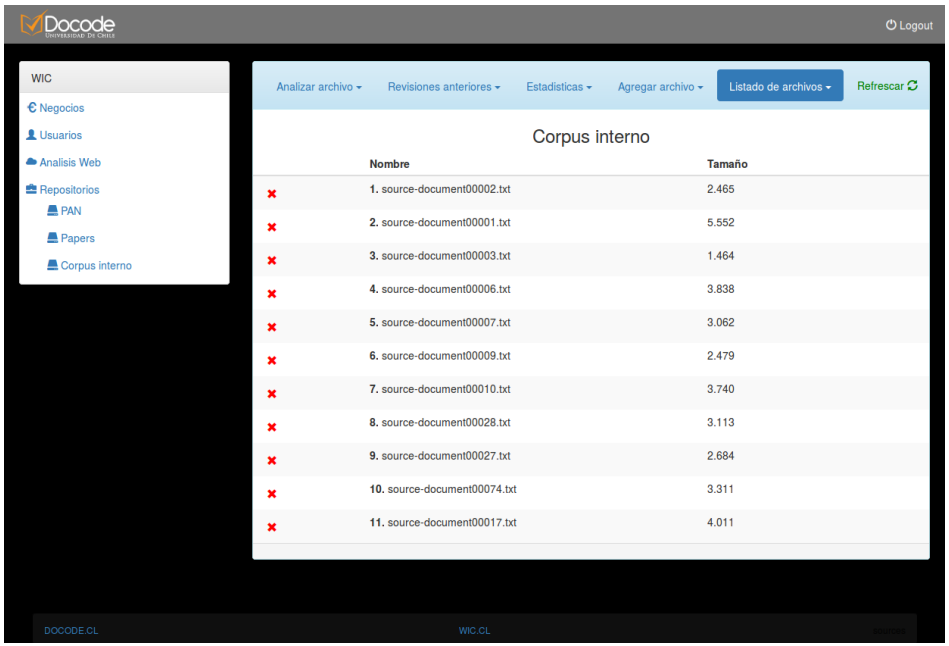
10. Conclusión
11. La colonia fue un periodo del tiempo amplio que abarco aproximadamente tres siglos, fue sencillo para el rey poder gobernar y controlar América, ya que solo traslado las costumbres y mandos de la monarquía española hasta América (Chile),el rey se quedo en España y mando gobernantes a América. En la colonia hubo educación, instituciones que hubo en ese tiempo (las primeras), economía los rangos sociales de ese tiempo, la sociedad colonial (jerárquicamente), la

Fuente

36. El primer colegio fundado en el país es el Seminario de La Imperial, en 1568. La primera escuela de la gramática se funda en 1591 en el convento de Santo Domingo, posteriormente en 1596 los Jesuitas abren otra escuela de gramática. Los Jesuitas y Dominicos siguieron impartiendo grados académicos y los colegios jesuitas recibieron el mayor caudal de alumnos. Los Jesuitas fundaron un internado para jóvenes aristócratas: el Convictorio de San Francisco Javier. Allí hizo sus primeros estudios Alonso de Ovalle y también el abate don Juan Ignacio Molina. Tras la expulsión de esa orden religiosa del país, en 1767, el internado, mantenido por el Estado, se transformó en el Convictorio Carolino. Las escuelas de gramática de los dominicos y jesuitas seguían a las escuelas primarias. Gozaban éstas de una categoría más elevada. Sus programas se centraban en el estudio de la gramática, filosofía, retórica y latín, que habilitaba para los estudios superiores.
37. Cada orden religiosa mantuvo estudios para la formación de los sacerdotes y se crearon los seminarios de Santiago y de Concepción.
38.
39. RÈGIMEN ESCOLAR EN LA ÈPOCA COLONIAL
40. Diferencias radicales separan las finalidades de la enseñanza colonial de la nuestra republicana. Siendo sacerdotes o, por excepción, maestros laicos, quienes tuvieron a su cargo la tuición de la juventud, consideraban el primer objetivo de su tarea la de formar el espíritu en el temor de Dios y la obediencia a la doble majestad: del sacerdocio y del rey. /.../ La finalidad cercana de los programas tendia a hacerlos converger hacia las disciplinas teològicas que abrazarian los jóvenes màs aprovechados. Los estudios estaban sujetos a la filosofía escolástica y el latín, y orientados hacia las ciencias sagradas. Es en este contexto que la enseñanza femenina no recibiera mayor atención. Tal como España, las colonias carecían de un sistema de educación popular. La monarquía jamás consideró un deber suyo la educación de las masas. La juzgaban una obra de caridad para órdenes religiosas como, por ejemplo, la de las Escuelas Pías, fundadas por el aragonés José de Calasanz, en 1617. En las colonias, los establecimientos no

Front-end full





Outline

The core algorithm

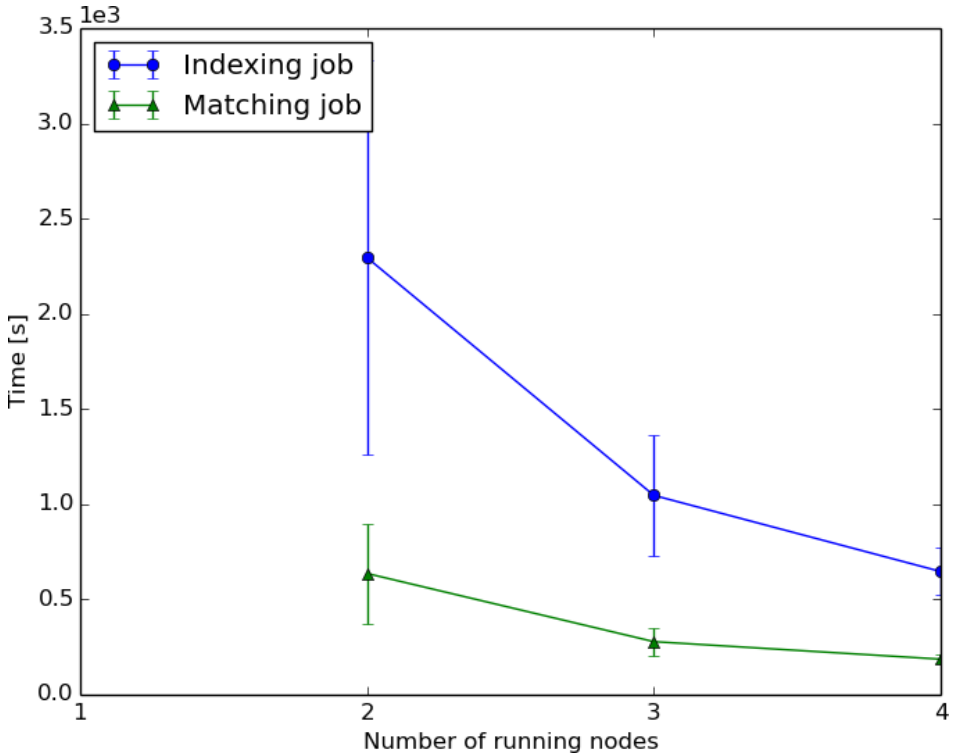
	Precision	**Recall**	**Granularity**	**Score**
Internal corpus	0.99	0.9	1.0	0.95
PAN-PC-13	0.88	0.99	1.04	0.9

Group analysis

- Hadoop cluster, 1 master, 4 slaves, 4GB RAM, 25 GB storage
- 110-MB dataset

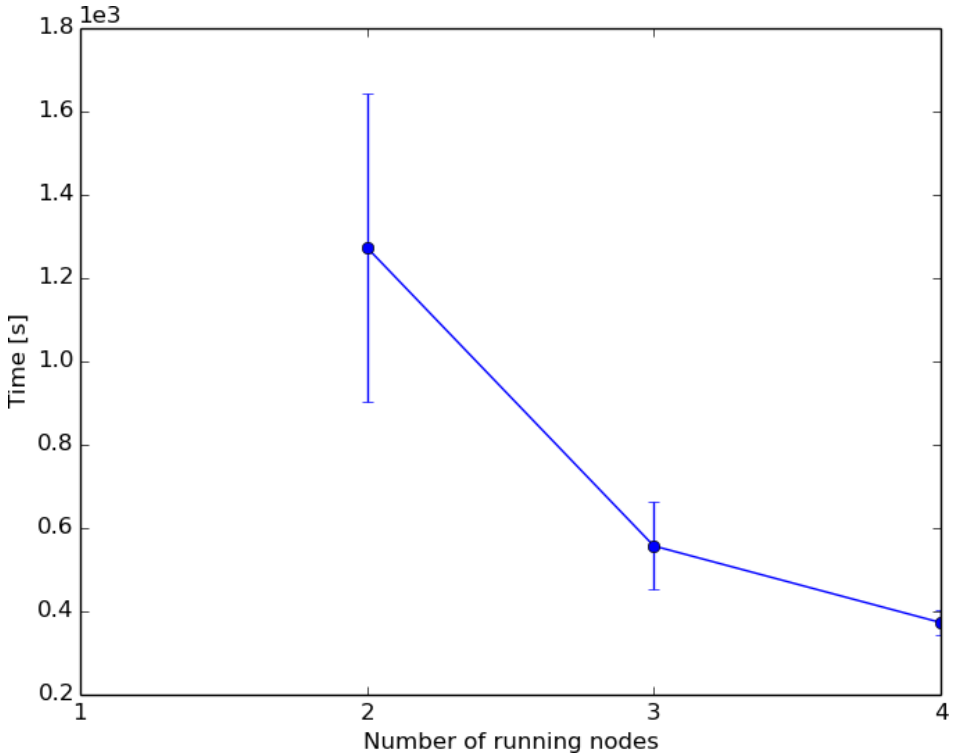
Group analysis

- Hadoop cluster, 1 master, 4 slaves, 4GB RAM, 25 GB storage
- 110-MB dataset



Group analysis

- Hadoop cluster, 1 master, 4 slaves, 4GB RAM, 25 GB storage
- 110-MB dataset



Outline

Conclusions

- Fully-fledged plagiarism detection engine
- Algorithms
- Deployment at scale

Conclusions

- Fully-fledged plagiarism detection engine
- Algorithms
- Deployment at scale

`http://docode.cl`

Future work

- Text extraction
- User interface
- Search engine

Docode 5: Building a scalable plagiarism detection system

Gaspar Pizarro V.

Web Intelligence Centre

November 22, 2016