

Uso de información de geolocalización, temporal y de usuario para el filtrado de información en Twitter



Víctor Fresno
vfresno@lsi.uned.es

Arkaitz Zubiaga
arkaitz@zubiaga.org
Heng Ji
hengjicuny@gmail.com



Natural Language Processing and
Information Retrieval Group at UNED

nlp.uned.es

Motivación

- ¿Qué es el filtrado de información?

Motivación

- ¿Qué es el filtrado de información?

“An Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human usefiltering problem”
(source: Wikipedia)

Motivación

- ¿Qué es el filtrado de información?

“An Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human usefiltering problem”

(source: Wikipedia)

- Nuestro **objetivo** es encontrar en Twitter toda la información relacionada con un evento específico durante un periodo corto de tiempo y en una zona geográfica específica.

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

"Un evento es algo que sucede en un momento determinado, en un lugar determinado y que puede involucrar o afectar a mucha gente".

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

"Un evento es algo que sucede en un momento determinado, en un lugar determinado y que puede involucrar o afectar a mucha gente".

Según la RAE: *"Suceso importante y programado, de índole social, académica o artística"*.

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

"Un evento es algo que sucede en un momento determinado, en un lugar determinado y que puede involucrar o afectar a mucha gente".

Según la RAE: *"Suceso importante y programado, de índole social, académica o artística"*.

Ejemplos: una conferencia científica, un huracán, una gala de premios, una manifestación como la del 25S, unas elecciones, ...

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

"Un evento es algo que sucede en un momento determinado, en un lugar determinado y que puede involucrar o afectar a mucha gente".

Según la RAE: *"Suceso importante y programado, de índole social, académica o artística".*

Ejemplos: una conferencia científica, **un huracán**, una gala de premios, una manifestación como la del 25S, unas elecciones, ...

Nosotros nos centramos en eventos relacionados con **"sucesos inesperados"**

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un **evento**?

"Un evento es algo que sucede en un momento determinado, en un lugar determinado y que puede involucrar o afectar a mucha gente".

Según la RAE: *"Suceso importante y programado, de índole social, académica o artística"*.

- ¿Son todos los eventos iguales?
 - ¿Generan la misma **cantidad de información**?
 - ¿Tienen el mismo impacto social?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar toda la información relacionada con un evento?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar **toda** la información relacionada con un evento?
 - ¿El acceso a la información en Twitter es igual que el que plantearíamos en Internet?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar **toda** la información relacionada con un evento?
 - ¿El acceso a la información en Twitter es igual que el que plantearíamos en Internet?
 - ¿Precision o Recall?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar **toda** la información relacionada con un evento?
 - ¿El acceso a la información en Twitter es igual que el que plantearíamos en Internet?
 - ¿Precision o Recall?
 - Información generada dinámicamente

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar **toda** la información relacionada con un evento?
 - ¿El acceso a la información en Twitter es igual que el que plantearíamos en Internet?
 - ¿Precision o Recall?
 - Información generada dinámicamente
 - Importante **componente temporal**: Información dependiente del instante.
 - Restringiendo el intervalo de tiempo modificamos la cantidad de información

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar toda la información relacionada con un evento?
- ¿Qué interés tiene el filtrado de información relativa a eventos en Twitter?

Motivación

- ¿Qué es el filtrado de información?
- ¿Qué es para nosotros un evento?
- ¿Tiene sentido tratar de encontrar toda la información relacionada con un evento?
- ¿Qué interés tiene el filtrado de información relativa a eventos en Twitter?
 - Encontrar subeventos, realizar análisis sociales, ...
 - Tarea previa a procesamientos de organización, descubrimiento de información, ...

Ejemplo real



Twitter 
@twitter

 Seguir

People sent more than 20 million Tweets about the storm between Oct 27 & Nov 1. Terms tracked: "sandy", "hurricane", **#sandy, #hurricane.**

 Responder  Retwittear  Favorito

561 RETWEETS	120 FAVORITES	
------------------------	-------------------------	--

9:46 am - 2 nov 12 por cpen · Insertar este Tweet

Ejemplo real



- Con este enfoque en nuestro dataset(*) obtenendríamos:
Precision: **0,970** Recall: **0,312** F1: **0,472**

(*) "Hurricane OR Isaac"

Ejemplo real



- Con este enfoque en nuestro dataset(*) obtenendríamos:
Precision: **0,970** Recall: **0,312** F1: **0,472**

(*) "Hurricane OR Isaac"

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?

“The idea behind relevance feedback is to take the results that are initially returned from a given query and to use information about whether or not those results are relevant to perform a new query. We can usefully distinguish between three types of feedback: explicit feedback, implicit feedback, and blind or "pseudo" feedback.” (source: Wikipedia)

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?
 - Divergencia de Modelos de Lenguaje: Kullback-Leibler Divergence (KLD)

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?
 - ¿Qué limitaciones presenta?

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?
 - ¿Qué limitaciones presenta?
 - Existencia de islas terminológicas !!

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?
 - ¿Qué limitaciones presenta?
 - ¿Cómo podemos superar esta limitación?

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
 - ¿Qué es el PRF?
 - ¿Cómo lo planteamos?
 - ¿Qué limitaciones presenta?
 - ¿Cómo podemos superar esta limitación?
 - Expandiendo a partir de tweets que NO contengan necesariamente los términos de consulta iniciales (*keyword/s* asociada/s inicialmente al evento)

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?
 - Información de geolocalización

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?
 - Información de geolocalización
 - Información de usuario

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?
 - Información de geolocalización
 - Información de usuario
 - Información temporal

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?
 - Información de geolocalización
 - Información de usuario
 - Información temporal
 - Hashtags

Planteamiento del problema

- Enfoque basado en expansión de consultas con pseudo-relevance feedback
- ¿Qué información podemos considerar en Twitter para la expansión de consultas?
 - Información de geolocalización
 - Información de usuario
 - Información temporal
 - Hashtags
 - Modificación de la función de expansión combinando las ideas del KLD y PRF

“Hurricane Isaac” dataset

- 1000 tweets geolocalizados publicados en el área de Louisiana el día 29 de Agosto de 2013

“Hurricane Isaac” dataset

- 1000 tweets geolocalizados publicados en el área de Louisiana el día 29 de Agosto de 2013
 - Timeframe: 8 a.m – 9 a.m.

“Hurricane Isaac” dataset

- 1000 tweets geolocalizados publicados en el área de Louisiana el día 29 de Agosto de 2013
 - Timeframe: 8 a.m – 9 a.m.
 - Área: se toman tweets en $\pm 2^\circ$ del ojo del huracán según información del US National Hurricane Center (<http://www.nhc.noaa.gov/archive/2012/ISAAC.shtml?>)

“Hurricane Isaac” dataset

- 1000 tweets geolocalizados publicados en el área de Louisiana el día 29 de Agosto de 2013
 - Timeframe: 8 a.m – 9 a.m.
 - Área: se toman tweets en $\pm 2^\circ$ del ojo del huracán según información del US National Hurricane Center (<http://www.nhc.noaa.gov/archive/2012/ISAAC.shtml?>)
 - Colección de referencia correspondiente a tweets del 19 al 28 de Agosto en el mismo área y timeframe.

“Hurricane Isaac” dataset

- 1000 tweets geolocalizados publicados en el área de Louisiana el día 29 de Agosto de 2013
 - Timeframe: 8 a.m – 9 a.m.
 - Área: se toman tweets en $\pm 2^\circ$ del ojo del huracán según información del US National Hurricane Center (<http://www.nhc.noaa.gov/archive/2012/ISAAC.shtml?>)
 - Colección de referencia correspondiente a tweets del 19 al 28 de Agosto en el mismo área y timeframe.
 - Preprocesado básico:
 - Stemming de Porter y
 - Eliminación de stopwords, tokens empezando con números o caracteres no alfanuméricos.

“Hurricane Isaac” dataset

- Tweets anotados como “related” o “non-related” siguiendo las siguientes guidelines:

“Hurricane Isaac” dataset

- Tweets anotados como “related” o “non-related” siguiendo las siguientes guidelines:
 - a) The tweet explicitly refers to the Hurricane Isaac.
 - b) The tweet refers to consequences of the Hurricane Isaac (e.g., power outage).
 - c) The tweet is aware of the hurricane, and would not have been sent otherwise (e.g., concerned about safety of friends).

Note: when a tweet does not provide evidence to be considered as related, it should be categorized as non-related.

“Hurricane Isaac” dataset

- Tweets anotados como “related” o “non-related” siguiendo las siguientes guidelines:
 - a) The tweet explicitly refers to the Hurricane Isaac.
 - b) The tweet refers to consequences of the Hurricane Isaac (e.g., power outage).
 - c) The tweet is aware of the hurricane, and would not have been sent otherwise (e.g., concerned about safety of friends).

Note: when a tweet does not provide evidence to be considered as related, it should be categorized as non-related.

- Mechanical Turk annotation task (5 anotaciones por tweet).

“Hurricane Isaac” dataset

- Tweets anotados como “related” o “non-related” siguiendo las siguientes guidelines:
 - a) The tweet explicitly refers to the Hurricane Isaac.
 - b) The tweet refers to consequences of the Hurricane Isaac (e.g., power outage).
 - c) The tweet is aware of the hurricane, and would not have been sent otherwise (e.g., concerned about safety of friends).

Note: when a tweet does not provide evidence to be considered as related, it should be categorized as non-related.

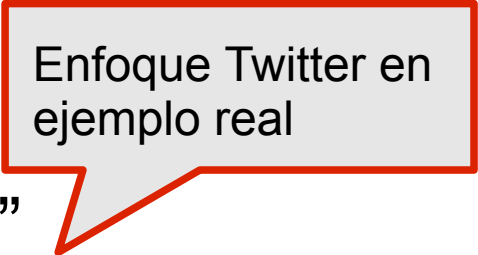
- Mechanical Turk annotation task (5 anotaciones por tweet).
- Resultado: 397 related and 603 non-related tweets

Experimentación (I)

- Baseline:
 - Query inicial (**Q**): “Hurricane OR Isaac”
 - Precision=0,985 Recall=0,3375

Experimentación (I)

- Baseline:
 - Query inicial (**Q**): “Hurricane OR Isaac”
 - Precision=0,985 Recall=0,3375



Enfoque Twitter en ejemplo real

Experimentación (I)

- Baseline:
 - Query inicial (**Q**): “Hurricane OR Isaac”
 - Precision=0,985 Recall=0,3375
- Baseline para la expansión:
 - Kullback-Leibler Divergence (KLD) aplicada a los tweets recuperados con Q (**Q+KLD**)

Experimentación (I)

- Baseline:

- Query inicial (**Q**): “Hurricane OR Isaac”

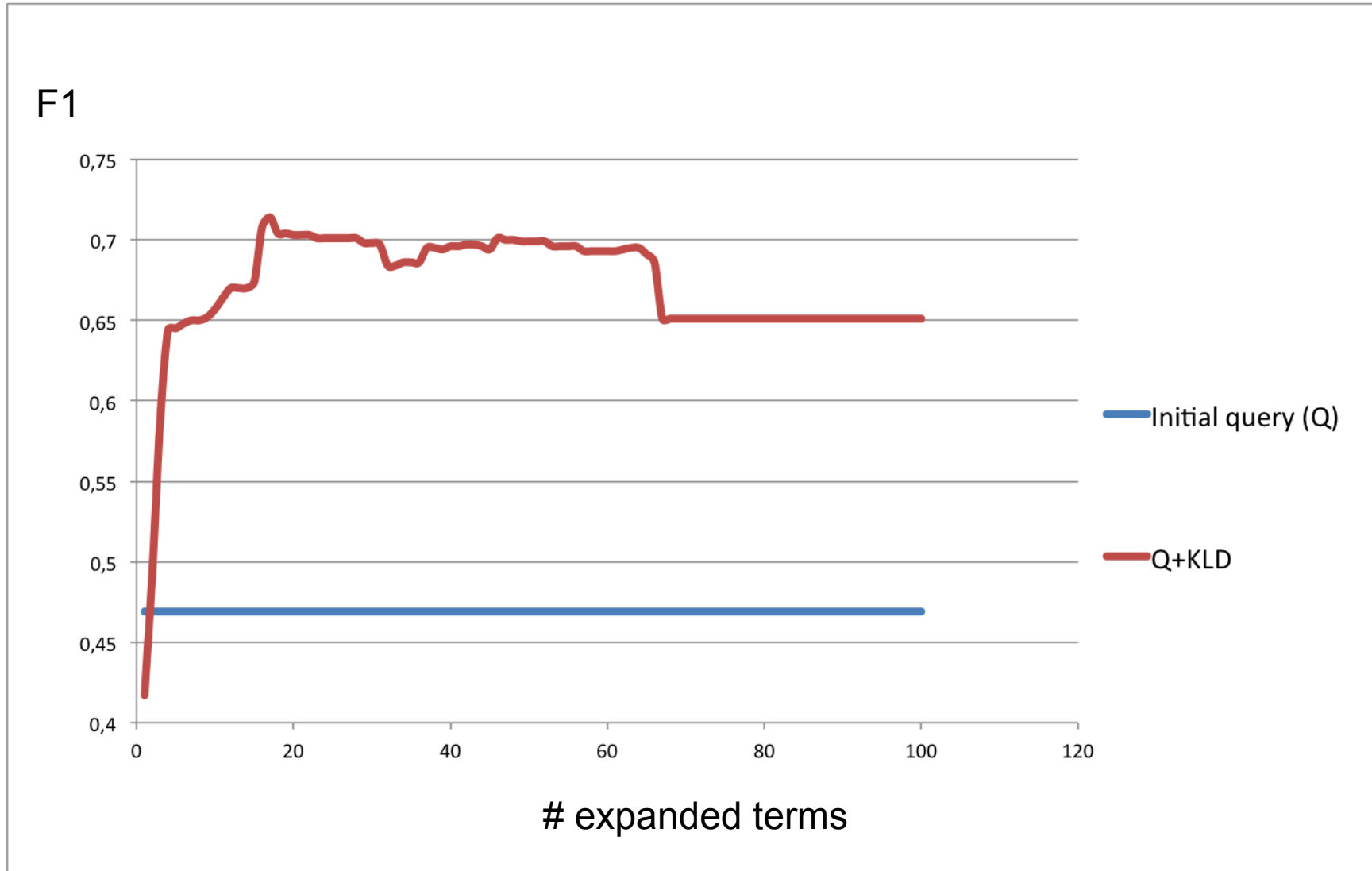
- Precision=0,985 Recall=0,3375

Típico enfoque de pseudo-relevance feedback

- Baseline para la expansión:

- Kullback-Leibler Divergence (KLD) aplicada a los tweets recuperados con Q (**Q+KLD**)

Resultados (baseline)



Experimentación (II)

- Proponemos una expansión considerando información de usuario:

Experimentación (II)

- Proponemos una expansión considerando información de usuario:
 - a) Expandimos con el tweet anterior de un usuario que twitteó “hurricane” o “isaac” ($Q+userPre+KLD$)

Experimentación (II)

- Proponemos una expansión considerando información de usuario:
 - a) Expandimos con el tweet anterior de un usuario que twitteó “hurricane” o “isaac” (Q+userPre+KLD)
 - b) Expandimos con el tweet siguiente de un usuario que twitteó “hurricane” o “isaac” (Q+userPost+KLD)

Experimentación (II)

- Proponemos una expansión considerando información de usuario:
 - a) Expandimos con el tweet anterior de un usuario que twitteó “hurricane” o “isaac” (Q+userPre+KLD)
 - b) Expandimos con el tweet siguiente de un usuario que twitteó “hurricane” o “isaac” (Q+userPost+KLD)
 - c) Expandimos con el tweet anterior y posterior de un usuario que twitteó “hurricane” o “isaac” (Q+userPrePost+KLD)

Experimentación (II)

- Proponemos una expansión considerando información de usuario:
 - a) Expandimos con el tweet anterior de un usuario que twitteó “hurricane” o “isaac” (**Q+userPre+KLD**)
 - b) Expandimos con el tweet siguiente de un usuario que twitteó “hurricane” o “isaac” (Q+userPost+KLD)
 - c) Expandimos con el tweet anterior y el posterior de un usuario que twitteó “hurricane” o “isaac” (Q+userPrePost+KLD)

Se encuentra que el mejor enfoque es a)

Experimentación (III)

- Proponemos una expansión considerando información de geolocalización:

Experimentación (III)

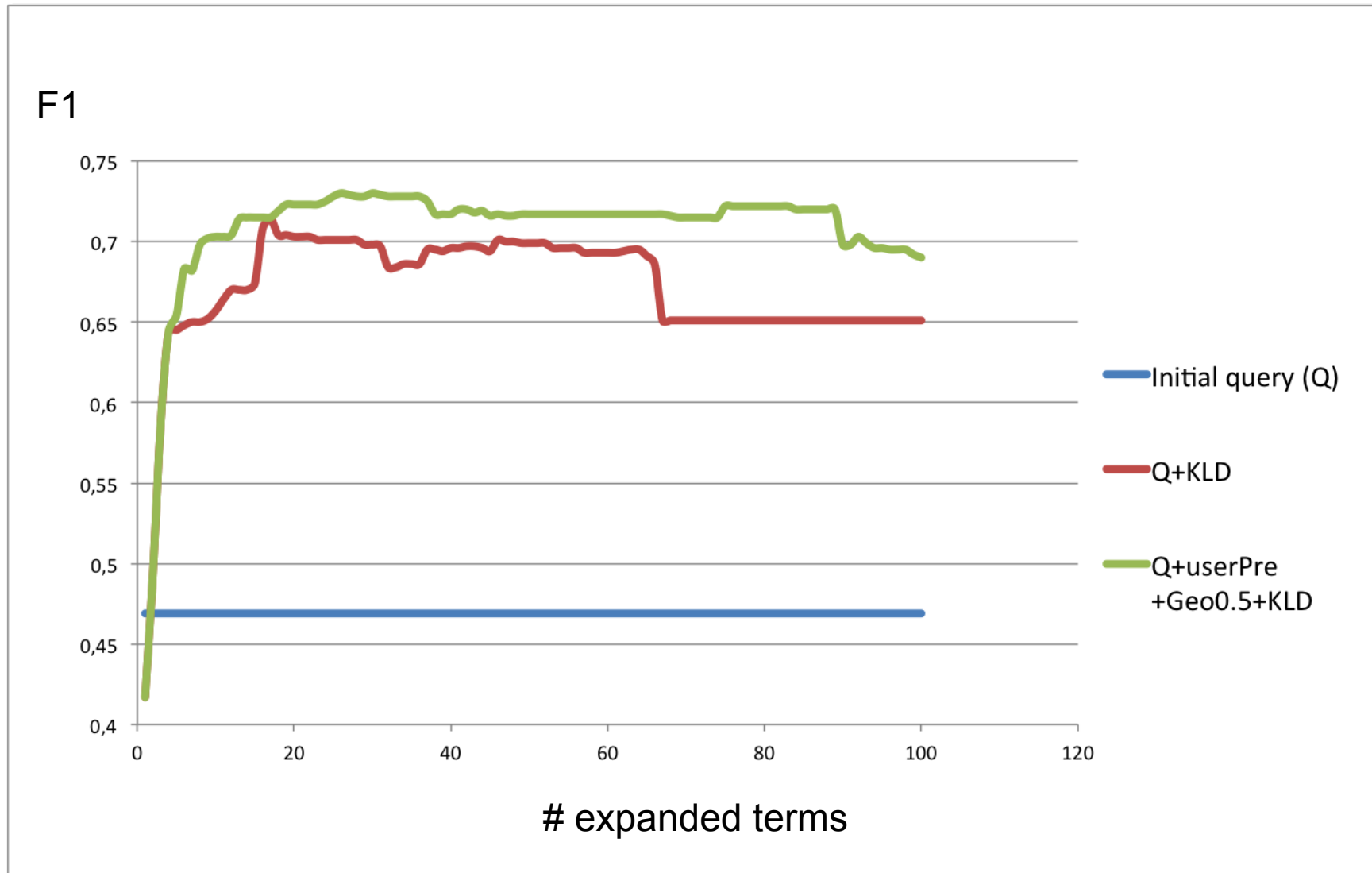
- Proponemos una expansión considerando información de geolocalización:
 - a) Expandimos con el tweet más cercano en un radio de $\pm 0.5^\circ$ de longitud y latitud para cada tweet que contuviera los términos “hurricane” o “Isaac” (Q+geo0.5+KLD)

Experimentación (III)

- Proponemos una expansión considerando información de geolocalización:
 - a) Expandimos con el tweet más cercano en un radio de $\pm 0.5^\circ$ de longitud y latitud para cada tweet que contuviera los términos “hurricane” o “Isaac” (Q+geo0.5+KLD)

Finalmente combinamos ambos enfoques
(Q+userPre+geo0.5+KLD)

Resultados (social information)



Experimentación (IV)

- Probamos considerar recursos lingüísticos externos (WordNet):

Experimentación (IV)

- Probamos considerar recursos lingüísticos externos (WordNet):
 - a) Expandimos considerando Q y todos los tweets que contuvieran NN y NNS en la glosa de “huracán” tras un proceso de POS tagging (Q+WN+KLD)

Experimentación (IV)

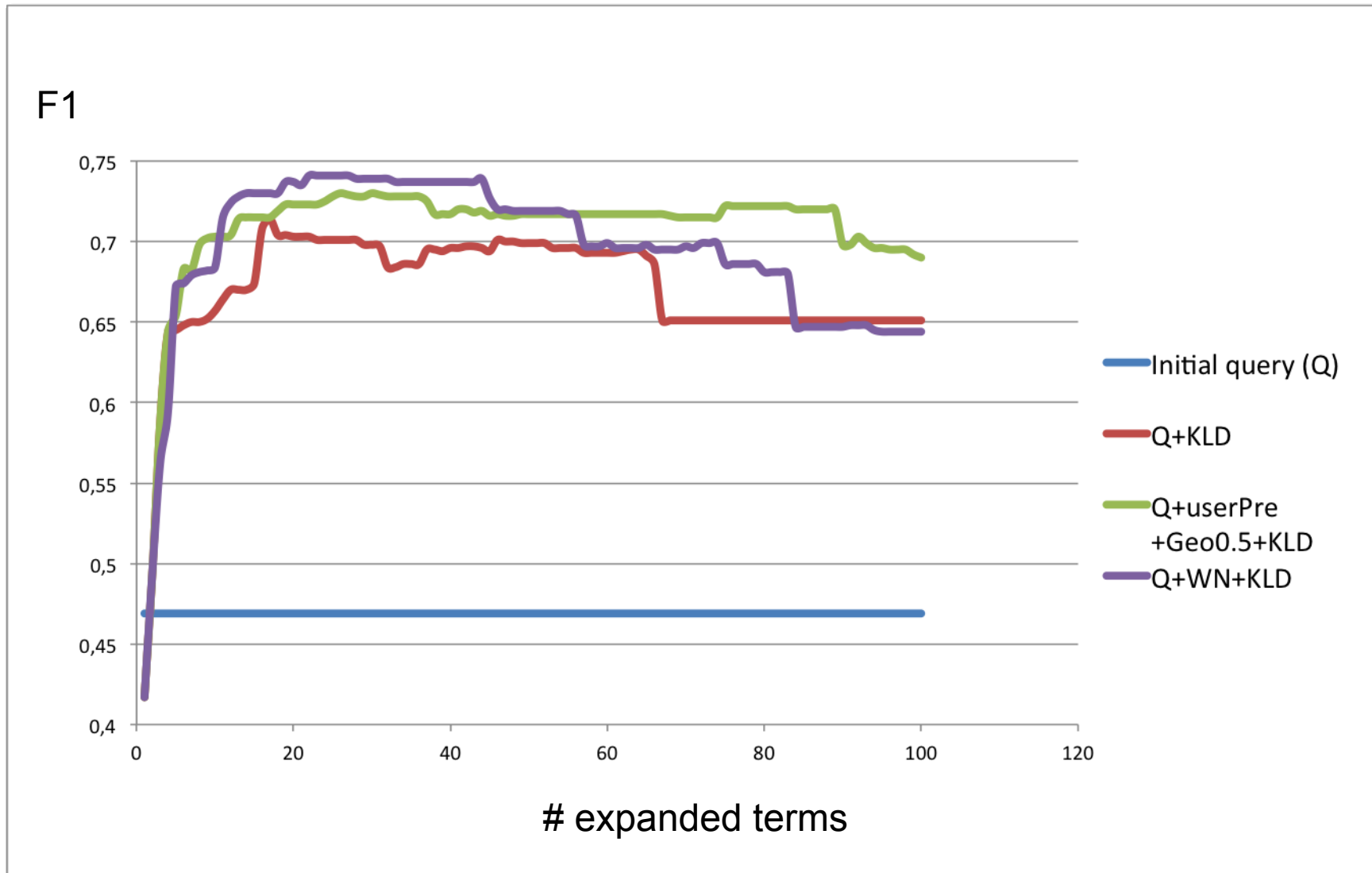
- Probamos considerar recursos lingüísticos externos (WordNet):

a) Expandimos considerando Q y todos los tweets que contuvieran NN y NNS en la glosa de “huracán” tras un proceso de POS tagging (Q+WN+KLD)

Hurricane: "a severe tropical cyclone usually with heavy rains and winds moving a 73-136 knots (12 on the Beaufort scale)".

DT/ a JJ/ severe JJ/ / tropical **NN/ cyclone** RB/ usually IN/
with JJ/ heavy **NNS/ rains** CC/ and **NNS/ winds** VBG/ moving
DT/ a CD/ 73-136 **NNS/ knots** -LRB-/ (CD/ 12 IN/ on DT/ the
NNP/ Beaufort NN/ scale -RRB-/)

Comparativa con WN



Experimentación (V)

- Probamos a modificar la función de expansión.

Experimentación (V)

- Probamos a modificar la función de expansión.

Baseline: Kullback Leibler divergence

$$KLD(t) = P(t|Q^*) \cdot \text{Log}[P(t|Q^*) / P(t|Aug10^{th}-Aug29^{th})]$$

being $P(t|Q^*)$ the probability of a give term t within a set of tweets $Q^* = \{Q, Q+userPre+Geo0.5, Q+WN\}$ and $P(t|Aug10^{th}-Aug29^{th})$ its probability within the whole Collection.

Experimentación (V)

- Proponemos una variante: Twitter-KLD*

Experimentación (V)

- Proponemos una variante: Twitter-KLD*

$$\begin{aligned} \text{Twitter-KLD}(t) = & \\ & P(t | \text{Aug29}^{\text{th}}) \cdot \\ & \text{Log} [P(t | \text{Aug29}^{\text{th}}) / P(t | \text{Aug10}^{\text{th}} - \text{Aug29}^{\text{th}})] \cdot \\ & \text{Log} [P(t | Q^*) / P(t | \text{Aug10}^{\text{th}} - \text{Aug29}^{\text{th}})] \end{aligned}$$

being $P(t | \text{Aug29}^{\text{th}})$ the probability of a given term t within tweets in Aug29^{th} and $Q^* = \{Q, Q + \text{userPre} + \text{Geo}0.5, Q + \text{WN}\}$

(*) Name suggestions are welcomed ;)

Experimentación (Λ)

- Proponemos una variante: Twitter

Divergencia de probabilidades entre un día y la colección completa (=KLD_day ~ TT)

$Twitter-KLD(t) =$

$P(t | Aug29^{th}) .$

$Log [P(t | Aug29^{th}) / P(t | Aug10^{th} - Aug29^{th})] .$

$Log [P(t | Q^*) / P(t | Aug10^{th} - Aug29^{th})]$

being $P(t | Aug29^{th})$ the probability of a given term t within tweets in $Aug29^{th}$ and $Q^* = \{Q, Q + userPre + Geo0.5, Q + WN\}$

(*) Name suggestions are welcomed ;)

Experimentación (V)

- Proponemos una variante: Twitter-KLD*

$$Twitter-KLD(t) = \frac{P(t|Aug29^{th})}{P(t|Aug10^{th}-Aug29^{th})}$$

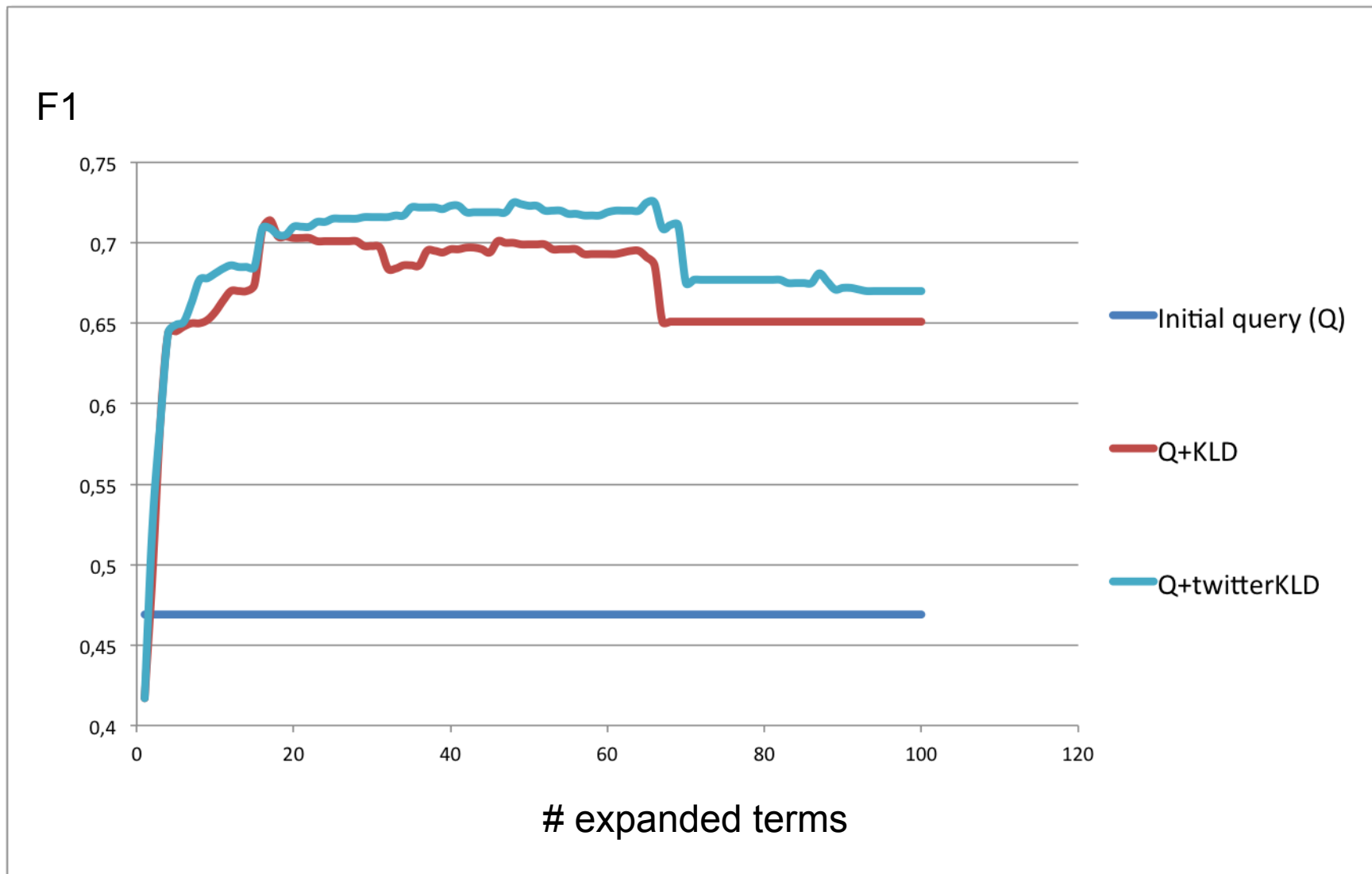
$$Log \left[\frac{P(t|Aug29^{th})}{P(t|Aug10^{th}-Aug29^{th})} \right]$$
$$Log \left[\frac{P(t|Q^*)}{P(t|Aug10^{th}-Aug29^{th})} \right]$$

Ratio entre la probabilidad de t en Q^* y en los días previos

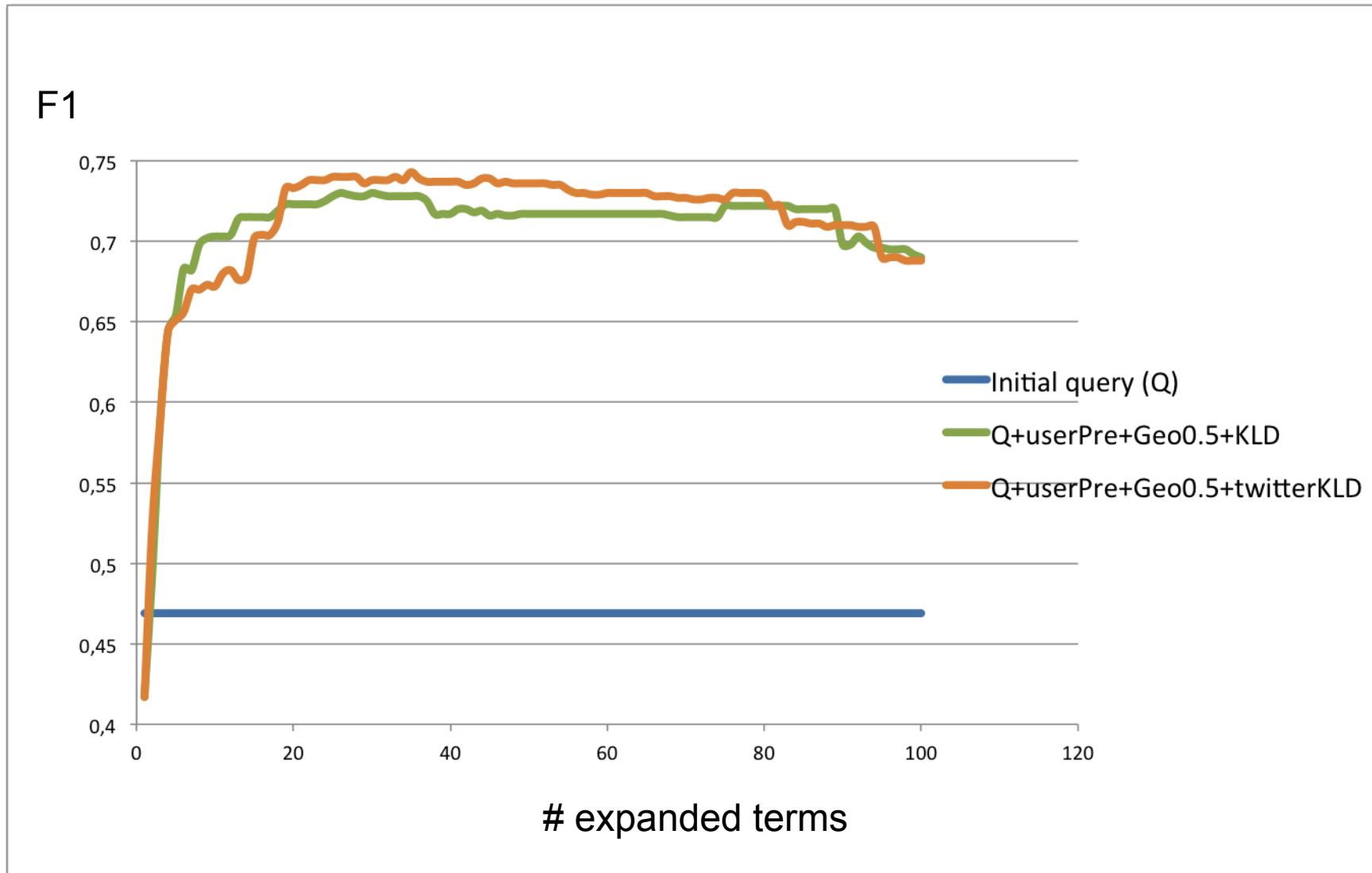
being $P(t|Aug29^{th})$ the probability of a given term t within tweets in $Aug29^{th}$ and $Q^* = \{Q, Q+userPre+Geo0.5, Q+WN\}$

(*) Name suggestions are welcomed ;)

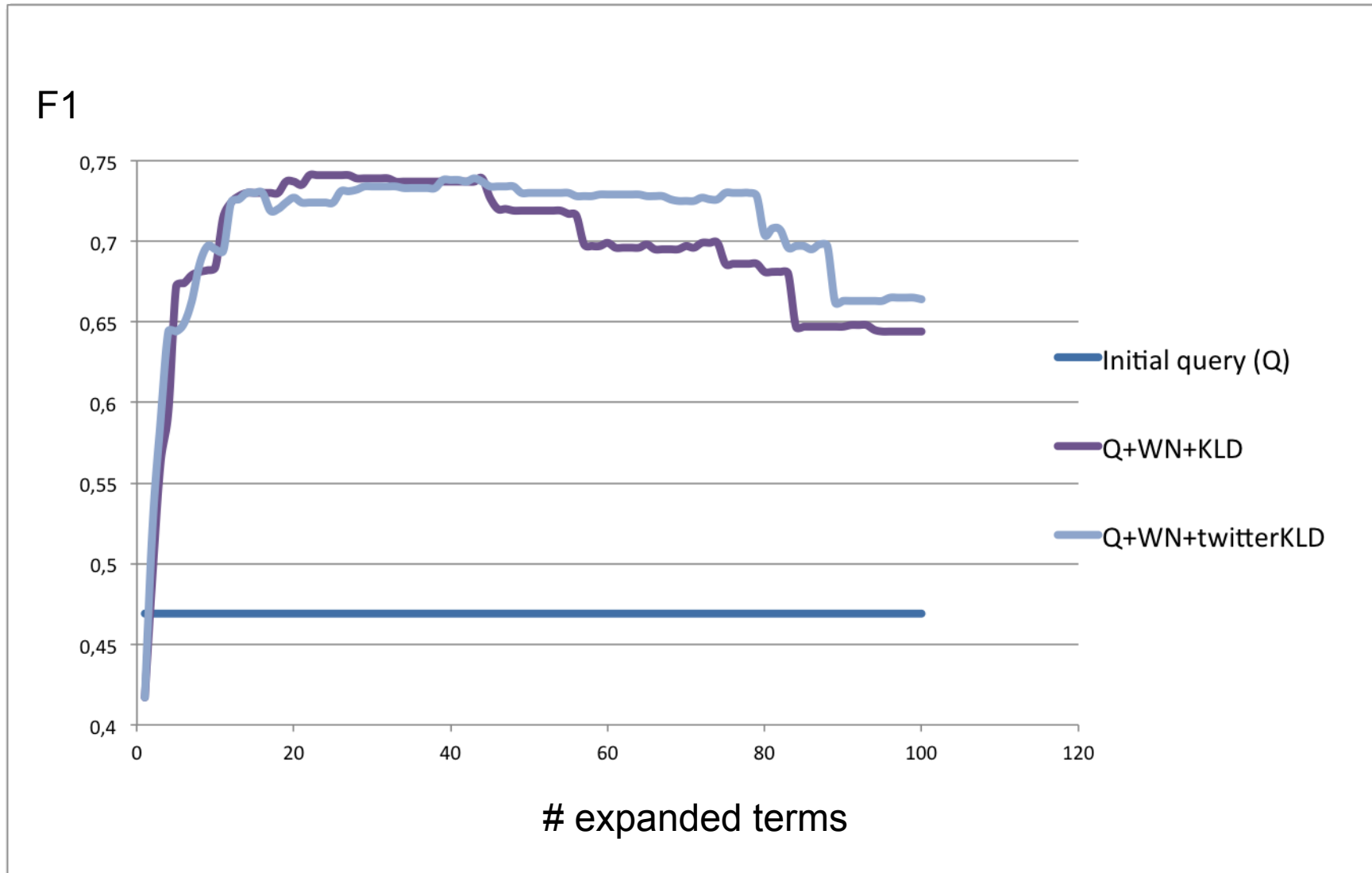
Evaluando *twitter-KLD* (I)



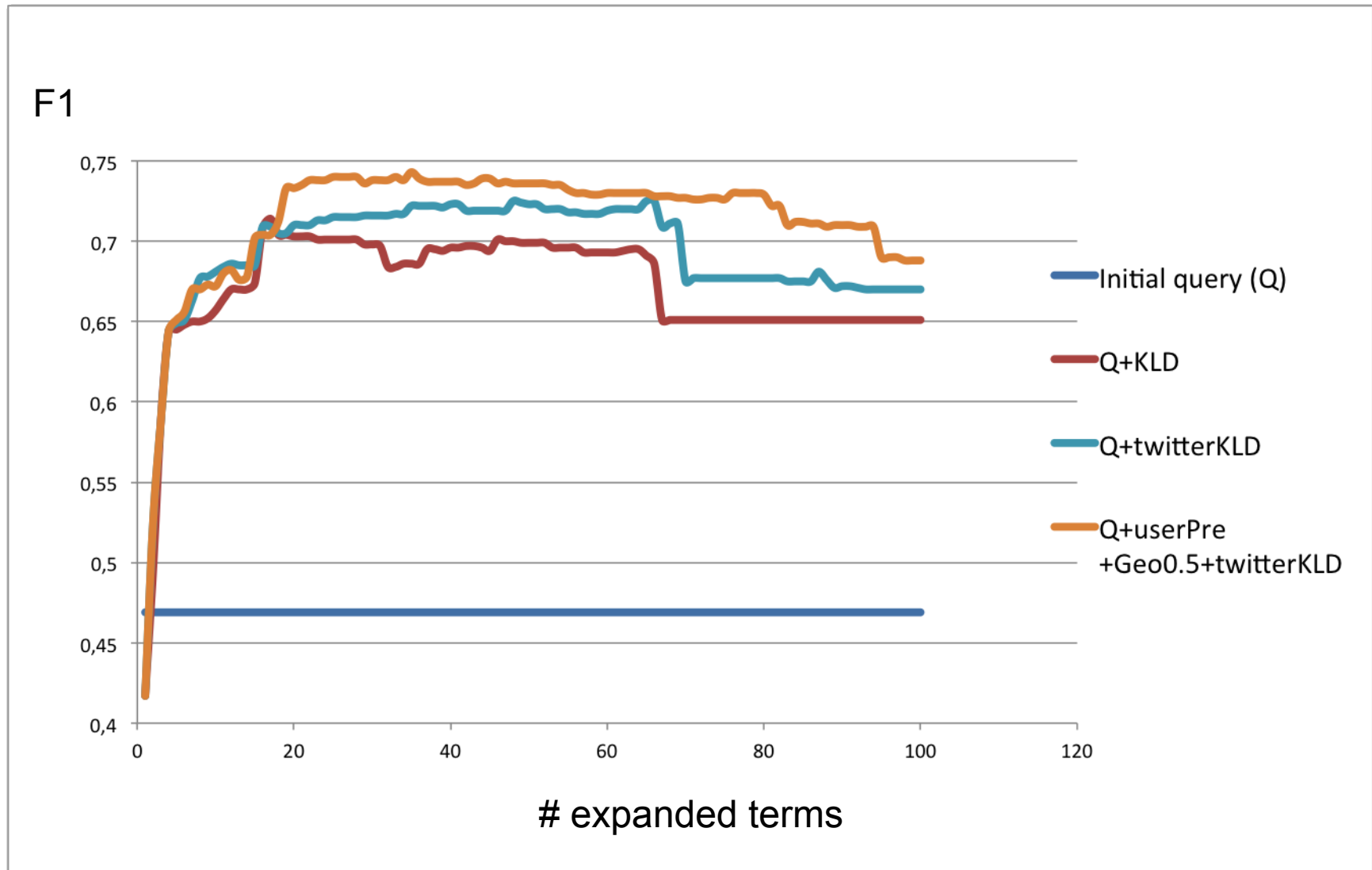
Evaluando *twitter-KLD* (II)



Evaluando *twitter-KLD* (III)



Evaluando *twitter-KLD* (IV)



Primeros análisis y conclusiones

- La limitación de la expansión por PRF produce “islas terminológicas” que puede superarse añadiendo “información social” (*información de usuario y geolocalización... hashtags, ...*)

Primeros análisis y conclusiones

- La limitación de la expansión por PRF produce “islas terminológicas” que puede superarse añadiendo “información social” (*información de usuario y geolocalización... hashtags, ...*)
- Podemos mejorar el filtrado de información relativa a un evento por medio expansión de consulta si consideramos info temporal en al función de expansión (TwKLD)

Primeros análisis y conclusiones

- La limitación de la expansión por PRF produce “islas terminológicas” que puede superarse añadiendo “información social” (*información de usuario y geolocalización... hashtags, ...*)
- Podemos mejorar el filtrado de información relativa a un evento por medio expansión de consulta si consideramos info temporal en al función de expansión (TwKLD)

Experimentos no mostrados aquí demuestran que el tipo de información social (*información de usuario y geolocalización... hashtags, ...*) a usar puede variar dependiendo de la colección (evento), pero que permite mejorar la tarea de filtrado.

Líneas futuras (I)

- Aplicación sobre diferentes datasets:
 - Hurricane Sandy
 - Colorado and Empire State shootings
 - 2012 US election in NYC and San Francisco

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas
 - Considerar agrupaciones de términos (Topic modelling? LSI?)

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas
 - Considerar agrupaciones de términos (Topic modelling? LSI?)
 - Extracción de conceptos de imágenes enlazadas (!!! :0 !!!)

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas
 - Considerar agrupaciones de términos (Topic modelling? LSI?)
 - Extracción de conceptos de imágenes enlazadas (!!! :0 !!!)
 - Aplicar segmentación en hashtags

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas
 - Considerar agrupaciones de términos (Topic modelling? LSI?)
 - Extracción de conceptos de imágenes enlazadas (!!! :0 !!!)
 - Aplicar segmentación en hashtags
 - Comparar la búsqueda por token y por subcadena (>3)

Líneas futuras (II)

- Posibles enfoques a explorar:
 - No considerar todos los tweets recuperados a partir de la expansión
 - Considerar menciones de @user y/o la autoridad
 - Considerar tweets recuperados con Q los días previos
 - Expandir no solo por el top k sino por el valor de relevancia
 - Expandir desde otras fuentes externas?
 - Considerar textos de URLs enlazadas
 - Considerar agrupaciones de términos (Topic modelling? LSI?)
 - Extracción de conceptos de imágenes enlazadas (!!! :0 !!!)
 - Aplicar segmentación en hashtags
 - Comparar la búsqueda por token y por subcadena (>3)
 - Considerar categoría gramatical para la expansión?
 - ...

¡¡ Gracias !!

Eskerrik asko !!

Thank you !!

谢谢

