Conceptualizing Documents with Wikipedia

Tadashi Nomoto
National Institute of Japanese Literature
nomoto@acm.org

Noriko Kando
National Institute of Informatics
kando@nii.ac.jp

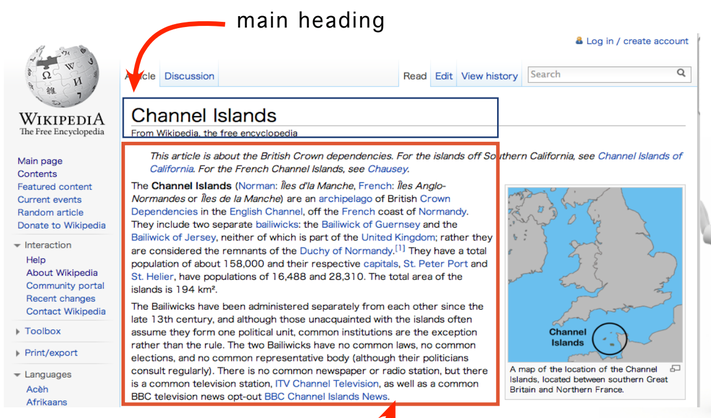
Goal

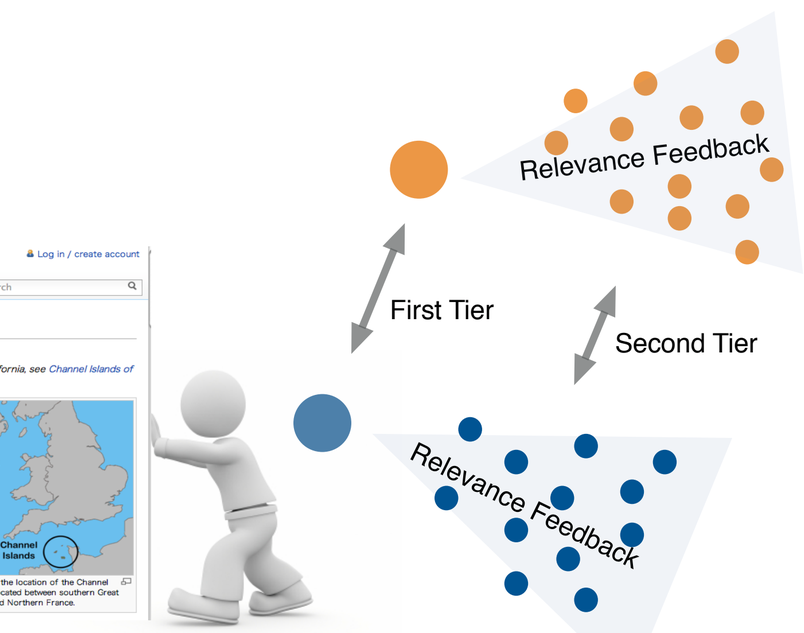
The goal is to find a way to label a document cluster with a natural, informative phrase, together with with some measure of confidence.

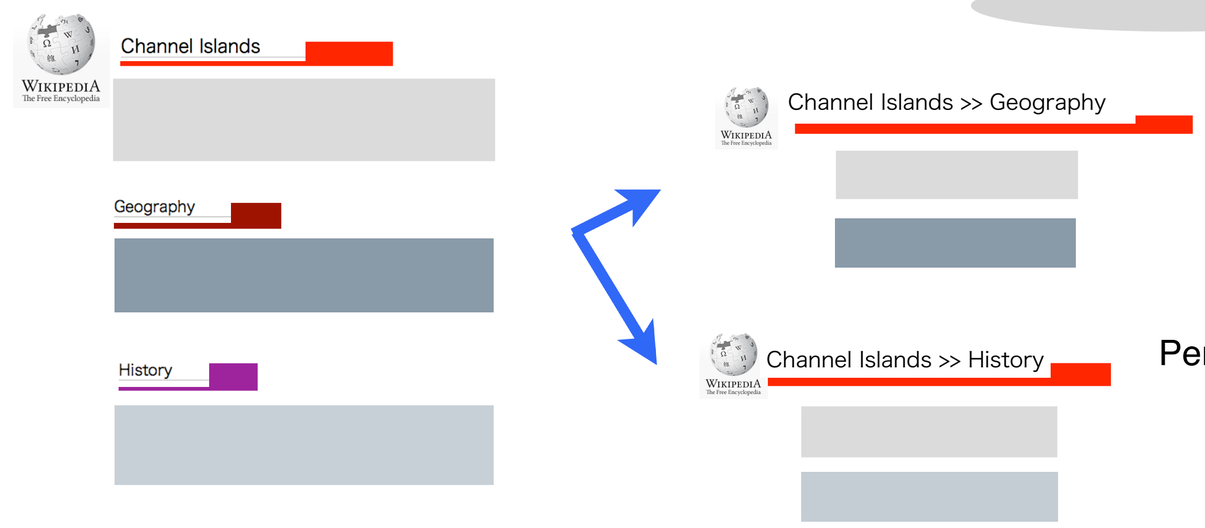
euro
debt
moody
greece
rating
europe
bailout
credit
crunch

European Sovereign Debt Crisis

Deconstructing Wikipedia







Similarity Measures

Cosine		$C(\mathbf{q},\mathbf{r})$	$=$	$\mathbf{q}^\top\mathbf{r}/\|\|\mathbf{q}\|\|_2\|\|\mathbf{r}\|\|_2$
L_1 norm	$(\|\|\mathbf{q}-\mathbf{r}\|\|_1^1)$	$L_1(\mathbf{q},\mathbf{r})$	$=$	$\sum_t \left\|\mathbf{q}(t)-\mathbf{r}(t)\right\|$
L_2 norm	$(\|\|\mathbf{q}-\mathbf{r}\|\|_2^2)$	$L_2(\mathbf{q},\mathbf{r})$	$=$	$\sum_t(\mathbf{q}(t)-\mathbf{r}(t))^2$
Polynomial Kernel		$P(\mathbf{q},\mathbf{r})$	$=$	$(\mathbf{q}^\top\mathbf{r}+C)^d$
RBF Kernel		$R(\mathbf{q},\mathbf{r})$	$=$	$\exp(-\sigma^2\|\|\mathbf{q}-\mathbf{r}\|\|_2^2)$
Hellinger		$H(\mathbf{q},\mathbf{r})$	$=$	$\sum_t\sqrt{\mathbf{q}(t)\mathbf{r}(t)}$
Skew divergence		$Q(\mathbf{q},\mathbf{r})$	$=$	$D(\mathbf{q}\|\|\,\alpha\,\mathbf{q}+(1-\alpha)\,\mathbf{r})$
Symmetric KL		$S(\mathbf{q},\mathbf{r})$	$=$	$D(\mathbf{q}\|\|\mathbf{r})+D(\mathbf{r}\|\|\mathbf{q})$

Two-Tiered Similarity Model (TTM)

$$\mathcal{S}(d_1,d_2)=\gamma\,sim_1(d_1,d_2)+(1-\gamma)\,sim_2(\uparrow d_1,\uparrow d_2)$$

Confidence Model (CFM)

$$P(R\geq t\mid\alpha_{\ell,d})=\frac{P^\star(\alpha_{\ell,d}\mid R\geq t)P^\dagger(R\geq t)}{P^\star(\alpha_{\ell,d})}$$

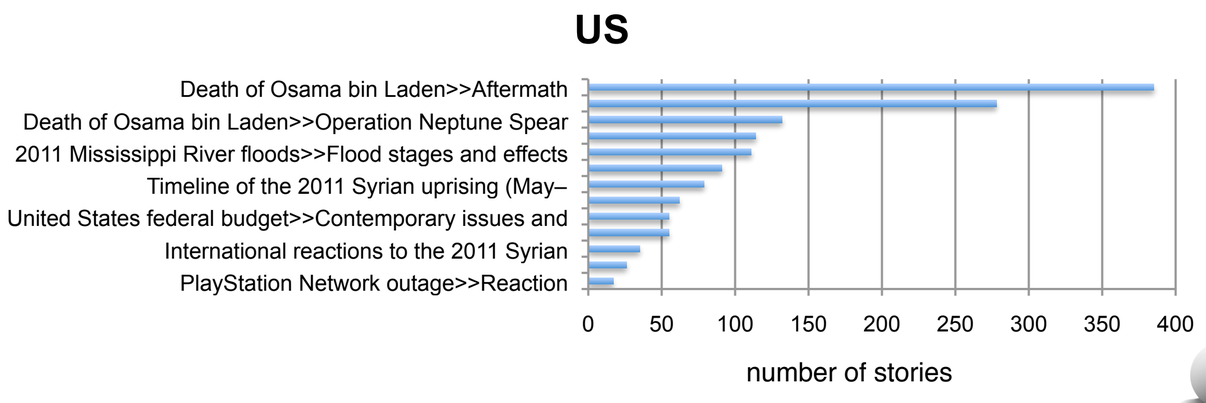
$$\pi(\ell,d;\mathcal{S})=\log P(R\geq t\mid\mathcal{S}(g(\ell),d))-\log P(R<t\mid\mathcal{S}(g(\ell),d))$$

Results

Performance in AUC with Wikipedia as a source for expansion at T2 (English)

MODEL	T1	T2	TTM	CFM(TTM)
C	0.678	0.665	0.693	**0.712**
L_1	0.684	0.664	0.689	**0.696**
L_2	0.673	0.694	0.707	**0.714**
P	**0.682**	0.612	0.632	0.677
H	0.682	0.612	0.631	**0.693**
Q	0.682	0.677	0.696	**0.709**
R	0.551	0.561	0.563	**0.633**
S	0.684	0.676	**0.689**	0.650
α	1.0	0.5	0.0	MAX
ELN	0.667	0.671	0.679	0.679

WikiLabel



$$l_\theta^*=\underset{l:p[l]\in\mathcal{W}^M}{\arg\max}\,Score(p[l],\theta|_N)$$

$$Score(p[l],\theta|_N)=\lambda\mathsf{COS}(p[l],\theta|_N)+(1-\lambda)\mathsf{OVL}(l,\theta|_N)$$