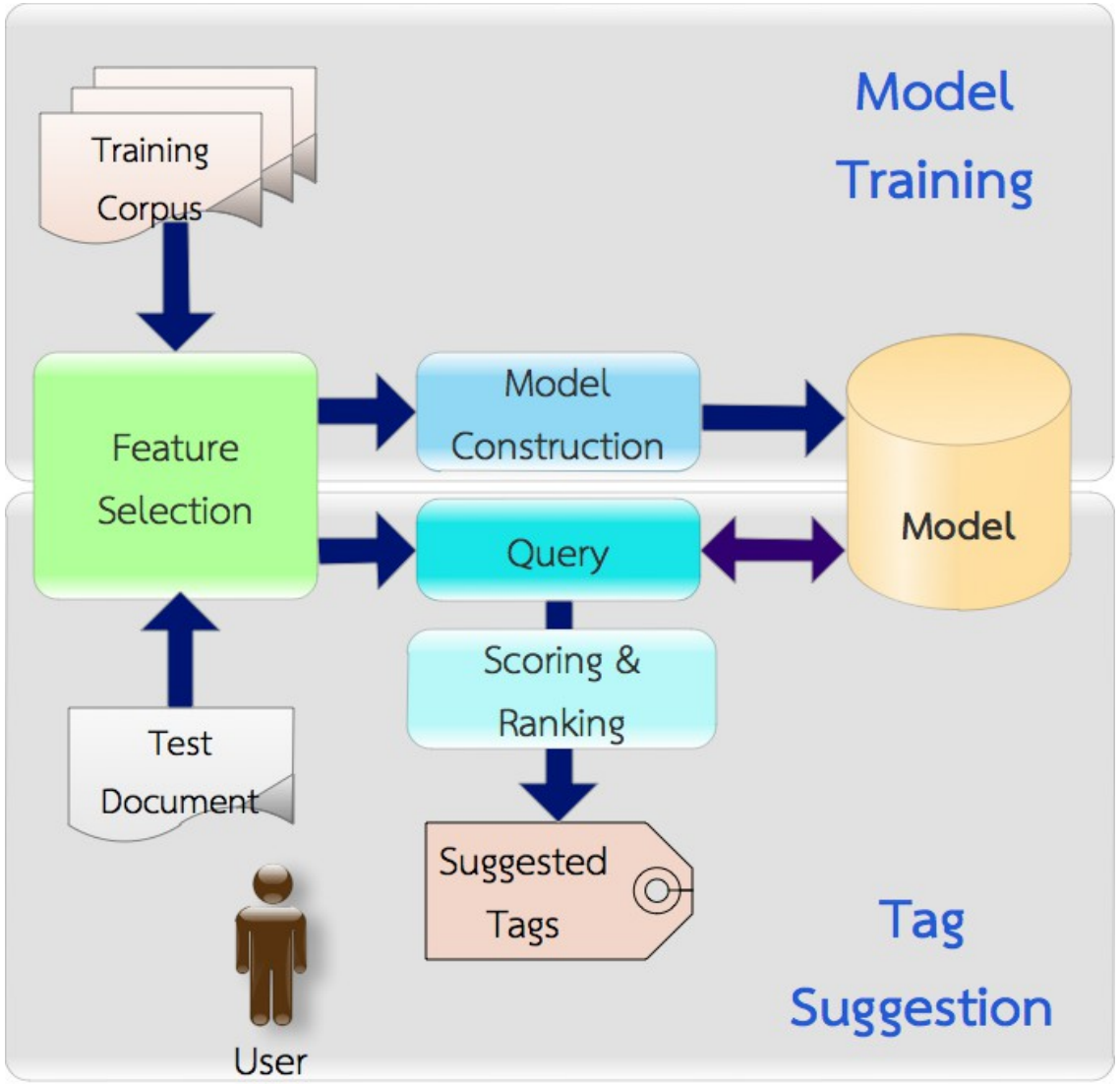
UREKA: Tag Suggestion

Framework overview of UREKA



UREKA: Tag

Corpus examples

Title: การจัดกิจกรรมเพื่อการท่องเที่ยวเชิงวัฒนธรรม :กรณีศึกษา ย่านท่าช้าง-ท่าพระจันทร์

Keywords: การท่องเที่ยวเชิงวัฒนธรรม , ไทย , กรุงเทพฯ , การศึกษาเฉพาะกรณี , การท่องเที่ยวเชิงวัฒนธรรม , ไทย , กรุงเทพฯ , การมีส่วนร่วมของพลเมือง

Title: การพิมพ์เอกสารอักษรเบรลล์จากโปรแกรมจัดพิมพ์เอกสารภาษาไทย/ภาษาอังกฤษ

Keywords: อักษรเบรลล์ , โปรแกรมคอมพิวเตอร์ , อักษรเบรลล์ , การประมวลผลข้อมูล

Title: ดารานักแสดงหญิงกับศัลยกรรมเสริมความงาม

Keywords: ความสวย , ศัลยกรรมตกแต่ง , นักแสดงสตรี , ไทย , การดำเนินชีวิต

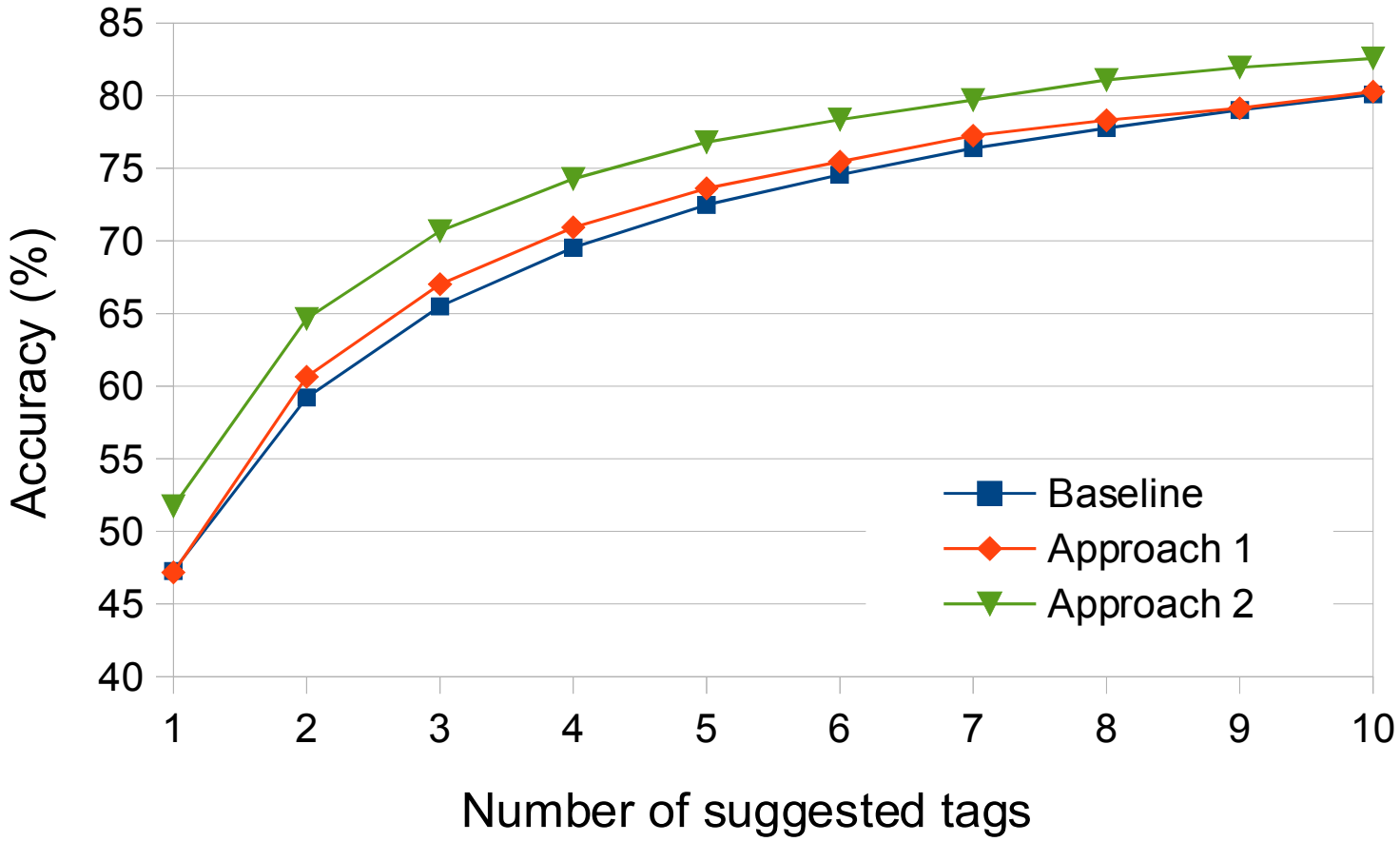
Title: การวิเคราะห์ตระกูลหนังผีไทย

Keywords: ภาพยนตร์เขย่าขวัญ , ไทย , ประวัติและวิจารณ์ , ผีในภาพยนตร์

Evaluation: Corpus stats

Statistics	
Total number of documents	14,481
Number of training documents	11,585
Number of test documents	2,896
Average number of word tokens /doc	14.24
Average number of tags /doc	4.18
Number of unique terms	10,222
Number of unique tags	6,372

Evaluation: results



Remark:

(1) Baseline approach: use all word tokens from title

(2) Approach 1: Stopword removal

(3) Approach 2: Stopword removal + append tags into title for training

Results discussion

Two challenging problems found in training data set

(1) Inconsistent tagging due to author subjectivity

Example:

Text: การศึกษาภาษาสนทนาทางโทรศัพท์ทางสถานีวิทยุข่าวสารและการจราจร (จส.100)

Ans: [การวิเคราะห์ภาษาระดับข้อความ, ภาษาไทย, บทสนทนาและวลี, การสื่อสารโดยการออกเสียง, ภาษาไทย, การวิเคราะห์ข้อความ]

Sys: [รายการวิทยุ, ผู้ฟังวิทยุ, การประเมินค่าความนิยม, การวิเคราะห์การสนทนา, กลุ่มสนทนาออนไลน์]

Results discussion

(2) Title contains specific word tokens which are not found

in training corpus

Example:

Text: การศึกษาเรื่อง "หนังประโมทัย" ในจังหวัดร้อยเอ็ด

Ans: [หนังตะลุง, การแสดงเงา, ไทย (ภาคตะวันออกเฉียงเหนือ)]

Sys: [ร้อยเอ็ด, อาชญากร, การคุมประพฤติ, อาสาสมัครในงานทัณฑวิทยา, การฟื้นฟูสมรรถภาพ]