

**EKSTRAKSI INFORMASI NAMA MAKANAN NUSANTARA  
BERDASARKAN ATURAN, KAMUS DAN SEGMENT-BASED  
HIDDEN MARKOV MODEL**

Tugas Akhir



Diajukan oleh:

WELLY PURNAMA

71130049

PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI  
UNIVERSITAS KRISTEN DUTA WACANA  
YOGYAKARTA  
2017

## PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

### **EKSTRAKSI INFORMASI NAMA MAKANAN NUSANTARA BERDASARKAN ATURAN, KAMUS DAN SEGMENT-BASED HIDDEN MARKOV MODEL**

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 31 Mei 2017



WELLY PURNAMA  
71130049

## HALAMAN PERSETUJUAN

Judul Skripsi : EKSTRAKSI INFORMASI NAMA MAKANAN  
NUSANTARA BERDASARKAN ATURAN,  
KAMUS DAN SEGMENT-BASED HIDDEN  
MARKOV MODEL

Nama Mahasiswa : WELLY PURNAMA

N I M : 71130049

Matakuliah : Skripsi (Tugas Akhir)

Kode : TIW276

Semester : Genap

Tahun Akademik : 2016/2017

Telah diperiksa dan disetujui di  
Yogyakarta,  
Pada tanggal 10 Mei 2017

Dosen Pembimbing I



Lucia Dwi Krisnawati, Dr.

Dosen Pembimbing II



Gloria Virginia, S.Kom., MAI, Ph.D.

## HALAMAN PENGESAHAN

### EKSTRAKSI INFORMASI NAMA MAKANAN NUSANTARA BERDASARKAN ATURAN, KAMUS DAN SEGMENT-BASED HIDDEN MARKOV MODEL

Oleh: WELLY PURNAMA / 71130049

Dipertahankan di depan Dewan Penguji Skripsi  
Program Studi Teknik Informatika Fakultas Teknologi Informasi  
Universitas Kristen Duta Wacana - Yogyakarta  
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar  
Sarjana Komputer  
pada tanggal 24 Mei 2017

Yogyakarta, 31 Mei 2017  
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr.
2. Gloria Virginia, S.Kom., MAI, Ph.D.
3. Sri Suwarno, Dr. Ir. M.Eng.
4. Yuan Lukito, S.Kom., M.Cs.


Dekan



(Budi Susanto, S.Kom., M.T.)



Ketua Program Studi



(Gloria Virginia, Ph.D.)

## UCAPAN TERIMAKASIH

Puji syukur dan terimakasih penulis panjatkan kepada Tuhan Yang Maha Esa atas segala berkat dan rahmat-Nya, sehingga laporan Tugas Akhir ini dapat diselesaikan.

Dalam penyusunan laporan ini tidak terlepas dari dukungan, bantuan, dan peran serta berbagai pihak yang telah membantu dalam proses penelitian dan pembuatan laporan tugas akhir ini. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Lucia Dwi Krisnawati selaku dosen pembimbing I.
2. Ibu Gloria Virginia, S.Kom., MAI, Ph.D. selaku dosen pembimbing II.
3. Kepada orang tua yang selalu memberi dukungan untuk menyelesaikan studi.

Laporan ini tentu tidak luput dari kesalahan dan kekurangan, mengingat keterbatasan kemampuan dan pengetahuan yang dimiliki penulis. Saran dan kritik yang bersifat membangun selalu penulis harapkan. Semoga laporan Tugas Akhir ini bermanfaat bagi penulis sendiri maupun bagi para pembaca.

Yogyakarta, ..... 2016

Welly Purnama

## **KATA PENGANTAR**

Puji syukur dan terimakasih penulis panjatkan kepada Tuhan Yang Maha Esa atas segala berkat dan rahmat-Nya, sehingga laporan tugas akhir ini dapat diselesaikan.

Laporan ini tentu tidak luput dari kesalahan dan kekurangan, mengingat keterbatasan kemampuan dan pengetahuan yang dimiliki penulis. Saran dan kritik yang bersifat membangun selalu penulis harapkan. Semoga laporan tugas akhir ini bermanfaat bagi penulis sendiri maupun bagi para pembaca.

Yogyakarta, ..... 2016

Welly Purnama

## INTI SARI

### EKSTRAKSI INFORMASI NAMA MAKANAN NUSANTARA BERDASARKAN ATURAN, KAMUS, DAN SEGMENT-BASED HIDDEN MARKOV MODEL

Banyaknya Informasi yang beredar bebas dan tidak terbatas melalui *website* maupun *social media* kini menimbulkan suatu masalah yang disebut dengan *Information Overload*. Salah satu penyebab informasi terbanyak di Internet adalah informasi mengenai nama makanan. Informasi tersebut mayoritas berhubungan dengan bisnis produksi, pembuatan, penjualan, dan iklan. Untuk mengatasi hal tersebut, maka dibutuhkan sistem yang dapat melakukan ekstraksi informasi. *Segment-based Hidden Markov Model* (SHMM) merupakan salah satu pemodelan statistika yang bisa diterapkan untuk membangun sistem ekstraksi informasi. Dalam penelitian ini, penulis akan menerapkan SHMM, leksikon, dan aturan untuk membangun sistem yang dapat melakukan ekstraksi nama makanan nusantara. Sistem juga akan dievaluasi dengan cara menghitung nilai presisi, *recall*, dan *f-measure*. Berdasarkan hasil analisis yang dilakukan, didapatkan bahwa penerapan SHMM, leksikon dan aturan untuk mengekstraksi nama makanan nusantara memiliki nilai *recall* yang tinggi, tetapi memiliki presisi yang rendah. Hal tersebut terjadi karena ditemukannya beberapa *redundancy* dari keluaran sistem. Selain itu, ditemukan juga bahwa dokumen latih memiliki peran yang sangat penting dalam ekstraksi nama makanan nusantara berdasarkan aturan, kamus, dan SHMM. Hal tersebut dibuktikan dengan nilai presisi dari dokumen uji yang diambil dari dokumen latih lebih tinggi daripada nilai presisi dari dokumen uji yang tidak diambil dari dokumen latih.

## DAFTAR ISI

HALAMAN JUDUL.....	
HALAMAN KEASLIAN SKRIPSI .....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PENGESAHAN.....	v
UCAPAN TERIMAKASIH.....	vi
KATA PENGANTAR .....	vii
INTI SARI.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR .....	xi
DAFTAR TABEL.....	xiv
BAB 1 PENDAHULUAN .....	1
1.1. Latar Belakang Masalah .....	1
1.2. Perumusan Masalah.....	2
1.3. Batasan Masalah .....	2
1.4. Tujuan Penelitian .....	3
1.5. Metode/Pendekatan .....	3
1.6. Sistematika Penulisan .....	4
BAB 2 LANDASAN TEORI.....	5
2.1. Tinjauan Pustaka .....	5
2.2. Landasan Teori .....	6
BAB 3 PERANCANGAN SISTEM .....	13
3.1. Pemilihan Bahasa Pemrograman .....	13
3.2. Spesifikasi Kebutuhan Sistem .....	13
3.3. Use Case .....	15



3.4.	Perancangan Proses.....	20
3.5.	Perancangan Basis Data.....	25
3.6.	Perancangan Antarmuka Sistem.....	28
3.7.	Perancangan Pengujian dan Evaluasi Sistem.....	30
BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM.....		32
4.2.	Implementasi Antarmuka.....	32
4.1.	Implementasi Segment-based Hidden Markov Model.....	39
4.3.	Implementasi Sistem dan Pseudocode.....	39
4.4.	Pengujian.....	43
4.5.	Analisis dan Diskusi.....	58
BAB 5 KESIMPULAN DAN SARAN.....		59
5.1.	Kesimpulan.....	59
5.2.	Saran.....	60
DAFTAR PUSTAKA.....		61
LAMPIRAN		

## DAFTAR GAMBAR

Gambar 2.1 Algoritma Viterbi mengambil jalur dengan probabilitas tertinggi....	10
Gambar 3.1 Diagram Use Case Sistem .....	15
Gambar 3.2 FlowChart pemrosesan Stopword .....	20
Gambar 3.3 FlowChart pemrosesan Leksikon .....	21
Gambar 3.4 FlowChart pemrosesan dokumen latih.....	22
Gambar 3.5 FlowChart pemrosesan dokumen uji.....	24
Gambar 3.6 FlowChart proses evaluasi .....	25
Gambar 3.7 Halaman Login.....	28
Gambar 3.8 Halaman menu admin dan pengguna .....	28
Gambar 3.9 Menu proses stopword .....	29
Gambar 3.10 Sub menu proses .....	29
Gambar 3.11 Sub menu list of .....	30
Gambar 4.1 Halaman login .....	33
Gambar 4.2 Halaman Home.....	33
Gambar 4.3 Halaman proses pada menu stopword.....	34
Gambar 4.4 Sub menu list of stopword.....	34
Gambar 4.5 Sub proses pada leksikon proses .....	35
Gambar 4.6 Sub menu list of leksikon .....	35
Gambar 4.7 Sub menu proses pada corpus proses .....	36

Gambar 4.8 Sub menu toke1 pada sub menu token .....	36
Gambar 4.9 Sub Menu Token2 pada sub menu token .....	37
Gambar 4.10 Sub menu proses pada IE-system.....	37
Gambar 4.11 Tampilan hasil ekstraksi.....	38
Gambar 4.12 Tampilah setelah tekan tombol detail.....	38
Gambar 4.13 Sub menu evaluasi pada menu IE-System .....	39
Gambar 4.14 Pseudocode proses stopwords.....	40
Gambar 4.15 Pseudocode proses leksikon .....	40
Gambar 4.16 Pseudocode mengolah token bigram.....	41
Gambar 4.17 Pseudocode mengolah token unigram.....	42
Gambar 4.18 Pseudocode untuk tokenisasi dan menghilangkan stopwords .....	43
Gambar 4.19 Pseudocode proses ekstraksi .....	43
Gambar 4.20 Hasil evaluasi Corpus A1 .....	44
Gambar 4.21 Grafik evaluasi Corpus A1 .....	45
Gambar 4.22 Hasil evaluasi Corpus A2.....	46
Gambar 4.23 Grafik evaluasi Corpus A2.....	47
Gambar 4.24 Hasil evaluasi Corpus A3.....	48
Gambar 4.25 Grafik evaluasi Corpus A3.....	50
Gambar 4.26 Hasil evaluasi Corpus B1 .....	50
Gambar 4.27 Grafik evaluasi Corpus B1 .....	50

Gambar 4.28 Hasil evaluasi Corpus B2 .....	52
Gambar 4.29 Grafik evaluasi Corpus B2 .....	53
Gambar 4.30 Hasil evaluasi Corpus B3 .....	54
Gambar 4.31 Grafik evaluasi Corpus B3 .....	56
Gambar 4.32 Perbandingan evaluasi dari 6 tahap .....	57
Gambar 4.33 Perbandingan evaluasi Corpus A dan Corpus B .....	57

©UKDW

## DAFTAR TABEL

Tabel 2.1 Tabel penerapan algoritma viterbi .....	10
Tabel 2.2 Tabel perhitungan unigram dengan add one smoothing .....	12
Tabel 3.1 Tabel Token .....	25
Tabel 3.2 Tabel Token2 .....	26
Tabel 3.3 Tabe nama_makanan.....	26
Tabel 3.4 Tabel stopwords.....	27
Tabel 3.5 Tabel evaluasi .....	27
Tabel 3.6 Tabel evaluasi2 .....	27
Tabel 4.1 Tabel hasil keluran corpus A1 .....	44
Tabel 4.2 Tabel hasil evaluasi corpus A2 .....	46
Tabel 4.3 Tabel hasil keluaran corpus A3.....	47
Tabel 4.4 Tabel hasil evaluasi corpus B1.....	51
Tabel 4.5 Tabel hasil evaluasi corpus B2.....	52
Tabel 4.6 Tabel hasil evaluasi corpus B3.....	55

## INTI SARI

### EKSTRAKSI INFORMASI NAMA MAKANAN NUSANTARA BERDASARKAN ATURAN, KAMUS, DAN SEGMENT-BASED HIDDEN MARKOV MODEL

Banyaknya Informasi yang beredar bebas dan tidak terbatas melalui *website* maupun *social media* kini menimbulkan suatu masalah yang disebut dengan *Information Overload*. Salah satu penyebab informasi terbanyak di Internet adalah informasi mengenai nama makanan. Informasi tersebut mayoritas berhubungan dengan bisnis produksi, pembuatan, penjualan, dan iklan. Untuk mengatasi hal tersebut, maka dibutuhkan sistem yang dapat melakukan ekstraksi informasi. *Segment-based Hidden Markov Model* (SHMM) merupakan salah satu pemodelan statistika yang bisa diterapkan untuk membangun sistem ekstraksi informasi. Dalam penelitian ini, penulis akan menerapkan SHMM, leksikon, dan aturan untuk membangun sistem yang dapat melakukan ekstraksi nama makanan nusantara. Sistem juga akan dievaluasi dengan cara menghitung nilai presisi, *recall*, dan *f-measure*. Berdasarkan hasil analisis yang dilakukan, didapatkan bahwa penerapan SHMM, leksikon dan aturan untuk mengekstraksi nama makanan nusantara memiliki nilai *recall* yang tinggi, tetapi memiliki presisi yang rendah. Hal tersebut terjadi karena ditemukannya beberapa *redundancy* dari keluaran sistem. Selain itu, ditemukan juga bahwa dokumen latih memiliki peran yang sangat penting dalam ekstraksi nama makanan nusantara berdasarkan aturan, kamus, dan SHMM. Hal tersebut dibuktikan dengan nilai presisi dari dokumen uji yang diambil dari dokumen latih lebih tinggi daripada nilai presisi dari dokumen uji yang tidak diambil dari dokumen latih.

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang Masalah

Makanan Indonesia merupakan salah satu makanan yang paling beragam di dunia. Kekayaan jenis makanannya merupakan cermin dari keberagaman budaya serta tradisi di Indonesia yang terdiri dari sekitar 6.000 pulau berpenghuni, dan menempati peran penting dalam budaya nasional Indonesia secara umum. (Day, 2015)

Menurut Ocepek dkk. (2013, hlm 1-5), informasi mengenai makanan adalah salah satu penyumbang informasi terbanyak di Internet melalui *website*, dan sosial media. Informasi tersebut mayoritas berhubungan dengan bisnis produksi, pembuatan, penjualan, dan iklan mengenai makanan. Krol (2011) berpendapat bahwa masalah yang saat ini dihadapi adalah *Information overload*, yaitu kondisi saat informasi beredar bebas dan tidak terbatas sehingga menyebabkan informasi yang diterima menjadi berlebihan.

Prado dan Freneda (2007: hlm 1-4) pada bukunya mengungkapkan bahwa banyaknya informasi pada volume halaman yang besar pada suatu dokumen membuat perlunya untuk mengekstrak informasi yang berguna dari dokumen. Ekstraksi informasi merupakan hal yang penting, dan sudah memberi efek yang besar sejauh untuk bidang penelitian, dan juga bisnis. Ekstraksi informasi nama makanan menjadi hal yang penting untuk mempermudah mendapatkan informasi yang lebih kompleks seperti meningkatkan mutu produk serta mengetahui produk yang laku di pasaran bagi pelaku bisnis dan penikmat makanan nusantara.

Oleh karena itu, untuk mengatasi masalah *information overload*, dan mempermudah para pelaku bisnis dan penikmat makanan dalam mencari informasi mengenai nama makanan nusantara, penulis akan membuat sistem cerdas untuk ekstraksi informasi nama makanan nusantara dari dokumen teks, dan membuang data-data tidak berguna agar *Information Overload* tidak terjadi pada pihak pelaku bisnis dan penikmat makanan nusantara. Sistem yang dibangun akan berbasiskan aturan, kamus, dan *Segment-Based Hidden Markov Model* dengan bahasa pemrograman PHP dan diimplementasikan dalam web.

## **1.2 Perumusan Masalah**

Dalam rangka membangun sistem dengan menggunakan aturan, kamus, dan *Segment-Based Hidden Markov Model*, ada beberapa pertanyaan yang harus dijawab :

1. Bagaimana cara mengimplementasikan aturan, kamus, dan *Segment-Based Hidden Markov Model* untuk ekstraksi nama-nama makanan nusantara ?
2. Bagaimana tingkat keakuratan dari penggunaan aturan, kamus dan *Segment-Based Hidden Markov Model* untuk mencari nama makanan nusantara pada suatu dokumen ?

## **1.3 Batasan Masalah**

Penelitian yang akan dilakukan memiliki beberapa batasan sebagai berikut ini :

1. Nama makanan nusantara yang dapat terekstraksi oleh sistem maksimal terdiri dari 3 kata, dan minimal 2 kata.
2. Sistem hanya bisa menerima masukkan file dokumen dalam bentuk .txt.



## 1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk membuat sebuah program yang mampu mengidentifikasi nama makanan nusantara dengan menggunakan aturan, kamus, dan *Segment-based Hidden Markov Model*. Sistem akan dievaluasi berdasarkan performa menggunakan perhitungan *precision*, *recall*, dan *F-measure*.

## 1.5 Metode Penelitian

Metode yang digunakan dalam penelitian ini antara lain sebagai berikut :

### 1. Pengumpulan data

Dokumen yang digunakan diambil dari artikel *online* (*blog*, *website*) yang berhubungan dengan sejarah, wisata, bisnis, dan hotel yang mengandung nama makanan nusantara.

### 2. Pembuatan aturan dan kamus

Aturan dan kamus yang di buat nantinya akan digunakan untuk proses akhir *filtering* nama makanan nusantara, selain itu kamus juga digunakan untuk proses awal mencari *segment*. Aturan dibuat oleh peneliti, sedangkan kamus didapatkan dari buku, artikel *online*, dan observasi secara langsung ke lapangan.

### 3. Pembangunan Sistem

Sistem akan dibuat dengan bahasa pemrograman PHP dan dikembangkan dengan aturan, kamus serta *Segment-Based Hidden Markov Model* dengan algoritma *viterbi* dan *smoothing* untuk mencari nama makanan nusantara dari suatu dokumen.

### 4. Evaluasi

Evaluasi bertujuan untuk mengukur performa sistem, dan akan dilakukan dengan cara menghitung nilai *precision*, *recall*, dan *F-measure* dari keluaran hasil ekstraksi sistem.

## 1.6 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini dibagi menjadi beberapa bab, yaitu:

Bab I Pendahuluan, berfungsi untuk memberikan gambaran umum tentang penelitian. Pendahuluan berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan laporan.

Bab II Tinjauan Pustaka, terdiri dari dua bagian utama, yaitu tinjauan pustaka dan landasan teori. Tinjauan pustaka menguraikan berbagai teori yang didapatkan dari berbagai sumber pustaka yang digunakan dalam penelitian. Sedangkan landasan teori memuat penjelasan tentang konsep dan prinsip utama yang diperlukan untuk memecahkan masalah.

Bab III Perancangan Sistem, berisi mengenai arsitektur sistem, spesifikasi sistem yang dibuat, diagram *use case*, algoritma dan *flowchart*

Bab IV Implementasi dan Analisis Sistem, berisi tentang implementasi dan analisis hasil akhir sistem secara keseluruhan dari pembuatan tugas akhir ini

Bab V Kesimpulan dan saran, berisi tentang kesimpulan dari hasil penelitian yang dilakukan serta saran untuk penelitian agar dapat memberikan hasil yang lebih baik.

## BAB V

### KESIMPULAN DAN SARAN

#### 5.1 Kesimpulan

Berdasarkan hasil implemmentasi dan analisis sistem, maka diperoleh kesimpulan sebagai berikut:

- Dari analisis evaluasi sistem, isi dan frekuensi dari dokumen uji dan dokumen latih sangat mempengaruhi hasil akhir sistem. Hal ini dikarenakan perhitungan probabilitas *Segment-based Hidden Markov Model* (SHMM) diambil dari dokumen uji dan dokumen latih.
- Dari hasil pengujian ditemukan bahwa penggunaan dokumen uji yang didapatkan dari dokumen latih (corpus A) memiliki nilai presisi yang lebih baik daripada dokumen uji yang didapatkan dari luar dokumen latih (corpus B), dan dari keseluruhan pengujian dokumen uji didapatkan nilai presisi yang rendah dengan nilai recall yang tinggi.
- Dari hasil analisis keluaran sistem, sistem tidak bisa melakukan ekstraksi terhadap nama makanan yang unik seperti oseng-oseng mercon, dan bakso setan karena keterbatasan jumlah leksikon dan variasi nama makanan pada dokumen latih.
- Nama kota bisa terekstraksi sebagai nama makanan jika terdapat pada leksikon dan dokumen latih seperti nasi padang, soto khas Boyolali, dll, sehingga dibutuhkan aturan khusus untuk nama makanan yang mengandung nama kota untuk meningkatkan presisi dari keluaran sistem.

## 5.2 Saran

Saran untuk pengembangan dan perbaikan sistem ini adalah :

- Diharapkan jumlah data latih yang digunakan diperbesar lagi.
- Isi dari dokumen latih diharapkan lebih disesuaikan lagi dengan topik nama makanan nusantara.
- Penambahan aturan baru yang khusus seperti aturan untuk nama makanan yang mengandung nama kota, seperti nasi Padang, soto Boyolali, dll.
- Penambahan aturan yang melibatkan konteks dimana nama makanan akan muncul atau prefiks leksikal untuk nama makanan.  
Contoh: kata “makan” berfungsi sebagai prefiks leksikal.
- Perluasan konteks perhitungan probabilitas dari *bigram* sampai *trigram* agar sistem dapat melakukan ekstraksi pada nama makanan nusantara yang lebih dari tiga kata.

## DAFTAR PUSTAKA

- Day, B. (2015, May 13). About Indonesian food. Diakses pada October 07, 2016, Dari <http://www.sbs.com.au/food/article/2008/07/01/about-indonesian-food>
- Gu, Zu., & Cercone, N. (2006, July). Segment-based Hidden Markov Models for Information Extraction. Proceedings of the 21<sup>st</sup> International Conference on Computational Linguistics and 44<sup>th</sup> Annual Meeting of the ACL, 481-488.
- Jurafsky, D S. (2000). Speech and Language Processing "An Introduction to Natural Language Processing, Cmputationak Linguistics, and Specch Reconition. Prentice-Hall, Inc. New Jersey.
- Kim, S., & Smyth, P. (2006, June 10). Segmental Hidden Markov Models with Random Effects for Waveform Modeling. *Journal of Machine Learning Research*,7 , 945-969.
- Krol, K. (2011, December 27). How to Fight Information Overload. Diakses pada October 9 .2016, dari <http://www.lifehack.org/articles/productivity/how-to-fight-information-overload.html>
- Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. New York:Cambridge University Press.
- Ocepek, M, G., Aspray, W., & Royer, G. (2013). *Food in the Internet Age*. Springer International Publishing.
- Prado, H. A., & Ferneda, E. (2007). *Emerging technologies of text mining: Techniques and applications*. Hershey, PA: Information Science.
- Piskorski, J., & Yangarber, R. (2012). Information Extraction: Past, Present and Future. *Multi-source, Multilingual Information Extraction and Summarization Theory and Applications of Natural Language Processing*, 23-49. doi:10.1007/978-3-642-28569-1\_2