Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA)

Authors:

Hee Jay Kang, Changhee Kim, Kyungtae Kang

Date Submitted: 2019-08-15

Keywords: LDA, research trend, topic modeling, biochemistry

Abstract:

Biochemistry has been broadly defined as “chemistry of molecules included or related to living systems”, but is becoming increasingly hard to be distinguished from other related fields. Targets of its studies evolve rapidly; some newly emerge, disappear, combine, or resurface themselves with a fresh viewpoint. Methodologies for biochemistry have been extremely diversified, thanks particularly to those adopted from molecular biology, synthetic chemistry, and biophysics. Therefore, this paper adopts topic modeling, a text mining technique, to identify the research topics in the field of biochemistry over the past twenty years and quantitatively analyze the changes in its trends. The results of the topic modeling analysis obtained through this study will provide a helpful tool for researchers, journal editors, publishers, and funding agencies to understand the connections among the diverse sub-fields in biochemical research and even see how the research topics branch out and integrate with other fields.

processes

Article

Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA)

Hee Jay Kang [1], Changhee Kim [1,*] and Kyungtae Kang [2,*]

[1] College of Business Administration, Incheon National University, 119, Academy-ro, Yeonsu-gu, Incheon 22012, Korea; hjkang86@snu.ac.kr

[2] Department of Applied Chemistry, Kyung Hee University, 1732, Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 130-701, Korea

[*] Correspondence: ckim@inu.ac.kr (C.K.); kkang@khu.ac.kr (K.K.); Tel.: +82-2-880-8594 (C.K.)

Received: 15 April 2019; Accepted: 13 June 2019; Published: 18 June 2019

Abstract: Biochemistry has been broadly defined as "chemistry of molecules included or related to living systems", but is becoming increasingly hard to be distinguished from other related fields. Targets of its studies evolve rapidly; some newly emerge, disappear, combine, or resurface themselves with a fresh viewpoint. Methodologies for biochemistry have been extremely diversified, thanks particularly to those adopted from molecular biology, synthetic chemistry, and biophysics. Therefore, this paper adopts topic modeling, a text mining technique, to identify the research topics in the field of biochemistry over the past twenty years and quantitatively analyze the changes in its trends. The results of the topic modeling analysis obtained through this study will provide a helpful tool for researchers, journal editors, publishers, and funding agencies to understand the connections among the diverse sub-fields in biochemical research and even see how the research topics branch out and integrate with other fields.

Keywords: biochemistry; topic modeling; research trend; LDA

1. Introduction

Biochemistry is the study of the structure, composition, and chemical reactions of substances in living systems and includes the sciences of molecular biology, immunochemistry, and neurochemistry, as well as bioinorganic, bioorganic, and biophysical chemistry [1].

Biochemistry has been broadly defined as "chemistry of molecules included or related to living systems", but is becoming increasingly hard to be distinguished from other related fields. Targets of its studies evolve rapidly; some newly emerge, disappear, combine, or resurface themselves with a fresh viewpoint. Methodologies for biochemistry have been extremely diversified, thanks particularly to those adopted from molecular biology, synthetic chemistry, and biophysics. There are sub-fields that are now regarded to lie within the field of biochemistry but used to be considered otherwise (e.g., nuclear magnetic resonance spectroscopy and mass spectroscopy for biological systems) [2–5].

Like other research fields these days, the field of biochemistry tends to focus on—and sometimes adjust itself to—a few high impactful themes, as shown by recent explosions of interest in gene-editing technologies, liquid–liquid phase separation, cryo-electron microscopy, or synthetic biology. Past topics similar to these, the so-called "hot" topics, however, sometimes were short-lived, as they matured rapidly or turned out to be less influential than they initially seemed to be. This is presumably because general access to research publication from the world-wide community is becoming easier and faster, so a few good papers published in powerful journals would give much more ramifications than they used to do [6–10].

The broad scope, rapidly changing interests, and fast transition of research topics in biochemistry make it an interesting field for trend studies. So far, the analysis of research trends has mostly been conducted using qualitative methodologies such as literature review, expert evaluations, and the Delphi method [11,12]. However, such qualitative techniques tend to require enormous time and costs to abstract significant results from large amounts of data while also carrying the possibility of bias depending on the scholars involved, as their subjective values or opinions may be reflected in the study. Moreover, a completely unbiased and objective evaluation of a field which encompasses a broad scope of research topics conducted over multiple decades can be a formidable task to even the top experts in the field. Particularly for biochemical research, whose trends are constantly shifting, a quantitative, as opposed to intuitive and popularity-driven, long-term trend analysis could provide a more objective and unbiased interpretation of the changes in research trends [13].

Therefore, this paper adopts topic modeling, a text mining technique, to identify the research topics in the field of biochemistry over the past twenty years and quantitatively analyze the changes in its trends. Topic modeling enables us to not only specify research topics so far touched upon by scholars in biochemistry but to also extract the keywords used in relation to the topics for a more in-depth analysis. Thus, the results of the topic modeling analysis obtained through this study will provide a helpful tool for researchers, journal editors, publishers, and funding agencies to understand the connections among the diverse sub-fields in biochemical research and even see how the research topics branch out and integrate with other fields. Also, for scholars and students of academic fields outside of biochemistry, this study will present an effective starting point for approaching biochemical research.

In Section 2 that follows, we summarize the existing literature that analyzes research trends using the topic modeling technique then explain the methods and application of the technique in Section 3. Section 4 describes the research data collected for this study and how it was preprocessed for our purposes. The analysis results are summarized in Section 5. Section 6 presents the conclusions of the research.

2. Literature Review

Topic modeling is an algorithm for locating topics from a large, unstructured collection of texts, and it is a model that infers topics by clustering words with similar meanings [14–16]. Because of this feature, topic modeling has been widely used to analyze topics and trends. Grimmer [17] analyzed the agendas of U.S. senators emphasized in their press releases, using topic modeling to examine how lawmakers inform their voters about their work. Mann et al. [18] demonstrated that topic modeling can be applied to measure the impact of research papers by applying topical n-grams (TNG) on 300,000 papers in the field of computer science.

Specific to the use of topic modeling to understand trends over time, Griffiths and Steyvers [15] used topic modeling to extract topics from abstracts listed on papers published between 1991 and 2001 in the proceedings of the National Academy of Sciences of the United States of America (PNAS), then identified the cold and hot topics by period [17]. Newman and Block [19] applied topic modeling to understand early American society and its publishing culture. Their study extracted topics from the text of newspapers published in the 18th century and analyzed how the topics changed over time. Gerrish and Blei [20] used the dynamic topic model to identify the changes in the topical contents over time in the corpus of academic research and to measure the influence of individual documents. Wang et al. [16] analyzed 17,000 studies published in Science, and Sun and Yin [13] analyzed the research trends in the transportation sector using topic modeling over time and by country.

As such, topic modeling is being applied to analyze existing literature or, notably, bibliographic data such as research abstracts, as a way to identify the research trends in diverse fields of study. Insofar as this present paper adopts the topic modeling technique on biochemical research, it may seem that what is attempted here lacks novelty. However, at the same time, the fact that topic modeling is frequently used in the existing literature across fields underscores the importance of identifying research trends. In particular, the blurry boundaries, the speedy evolution of topics, and the openness to convergence studies which are characteristic of the field of biochemistry reinforces the contributions of this present study.

3. Topic Modeling

Topic modeling is a text mining technique for discovering an abstract 'subject' from a set of documents. A document is generally written on one topic, and as such, the words related to the topic would appear more often than the other words in the document. For example, in a document on the subject of dogs, the words "dog" and "bones" would appear more often, while it is assumed that a document on the topic of cats will more often contain the words "cat" and "meow." A topic model, roughly speaking, binds the words "dogs" and "bones" under one topic, and "cat" and "meow" under another topic. Topic modeling, like the K-means clustering technique, sets the number of topics in advance and endows the subjects to the words grouped under topics at a later stage.

Latent Dirichlet allocation (LDA), a representative topic modeling technique, is a model based on procedural probability distribution that finds potentially meaningful topics in multiple documents [21]. LDA analysis calculates the probability that certain words will be included in each topic, assuming that multiple words can be grouped under different topics, and calculates the probability that those words will be included in each topic to extract a set of words with high probabilities corresponding to a topic. That is, LDA analysis finds the latent topic corresponding to the words in any given document. The schematic of the LDA technique can be visualized as Figure 1.

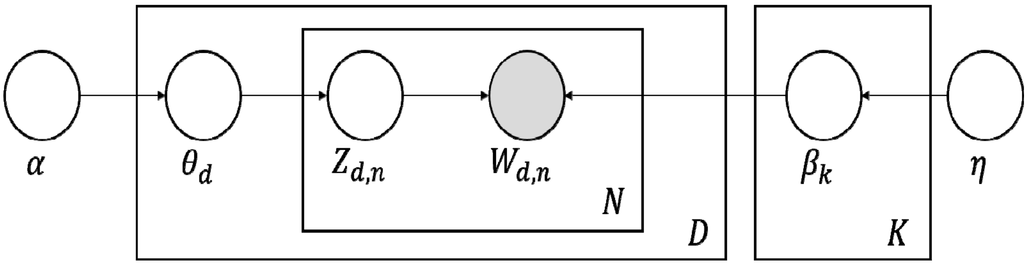

Figure 1. Schematic of the topic modeling algorithm. K: Number of topics; α: Dirichlet prior weight of topic k by document, the parameter which determines the value of θ; η: Dirichlet prior weight of word w by document, the parameter which determines the value of β; θ_d: The ratio of topics by document; β_k: The probability that word w will be generated by topic; $Z_{d,n}$: The topic of the nth word in document d (index); $W_{d,n}$: The nth word in document d (variable observed in document, index).

LDA's algorithm finds the latent subject of a document by inferring a hidden variable based on the variables observed in the document, where the observed variables are words ($W_{d,n}$). The algorithm uses hyper parameters α and η and the hidden parameter β_k to extract the words. The hidden variables $Z_{d,n}$ and $W_{d,n}$ cannot be observed directly in the document but can be inferred through the LDA model. In the LDA model, $Z_{d,n}$ is generated from θ_d, which is the ratio of topics by document whose value follows the Dirichlet prior weight determined by the value α. Likewise, β_k, which is the probability that a word will be generated by topic, is determined by the value η, and the Dirichlet prior weight of β_k is shaped by η. The word $W_{d,n}$ is thus identified by $Z_{d,n}$, the value that shows the topic of each word, and β_k, the word-by-topic ratio. The algorithm can be expressed as an equation, as follows:

$$p(z_1, \cdots, z_N) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n|\theta)\right)d\theta \tag{1}$$

$$p(w, z) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n)\right)d\theta \tag{2}$$

$$p(\theta,\, z\big|w,\, \alpha,\, \beta) = \frac{p(\theta,\, z,\, w\big|\alpha,\, \beta)}{p(w\big|\alpha,\, \beta)} \tag{3}$$

Other than LDA, there are also topic modeling algorithms such as latent semantic analysis (LSA), probabilitic LSA (pLSA), and Dirichlet multinomial regression (DMR). LSA builds semantic spaces based on a corpus and compares the similarities between words, sentences, paragraphs, and documents to form word clusters [22–25]. While LSA measures the similarities and creates clusters based on the frequency at which words are used in a document, pLSA is a model which looks at the probability that a specific word will appear in a document. LDA is a Bayesian version of pLSA using the Dirichlet distribution, which is a conjugate prior. Thus, while pLSA uses only the document-term matrix as input without consideration of the distribution of topics in a document, LDA considers both the distribution of topics by document and the distribution of terms by topic [26]. DMR expands on LSA to assume that the hyper parameter α depends on the document's metadata (author, year, department, country) [27,28].

At present, LDA is the most popular topic modeling algorithm used by scholars. Several studies compare LSA and LDA, but there has been no definite conclusion on whether one method is dominantly superior to the other [27,29]. Because LSA is based on term frequency, its advantage is that it produces intuitive results. On the other hand, the strength of LDA is that, because it is a probability-based model, it can reveal hidden connections which cannot be found by looking only at frequency. As this study attempts to redefine the topics in biochemistry and analyze their trends, we utilized the LDA model. Also, because it is difficult to specify the metadata of bibliographical information (e.g., the 'geographical area' of a paper can be difficult to define when there are more than one authors who are of different affiliations), the DMR model was considered unsuitable for our research.

4. Data Collection and Preprocessing

Among the research papers provided by the American Chemical Society from 1999 to 2018, this study analyzed 52 journals and 26,422 biochemical papers that fall under the subject of "general chemistry" on the American Chemical Society's research database (ACS Publications https://pubs.acs.org/). The amount of data collected by journal and year are given in Table 1. The journals Biochemistry, Journal of Physical Chemistry B, Journal of the American Chemical Society, and Langmuir, which published the largest number of relevant papers, have the impact factors of 2.938, 3.146, 14.357, and 3.789, respectively. As can be seen from Figure 2, there has been a steady increase in the number of published papers on biochemistry from 1999, with the largest number of studies published in 2012, followed by a slight decline (It should be mentioned that the significant drop in the number of papers published in 2018 is because the ACS database has not been fully updated after June 2018, unrelated to the trends in biochemistry research (Figure 2). Only the articles clearly categorized as biochemistry research in the ACS database were used as the primary data for this study).

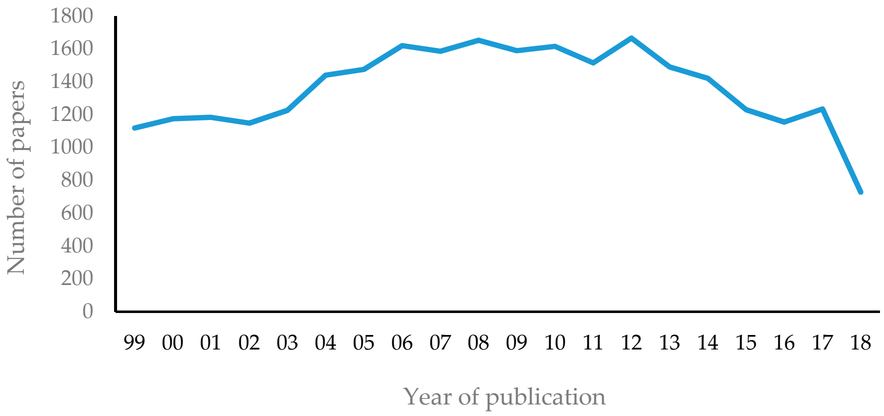

Figure 2. Number of biochemical papers published by year.

Table 1. Number of biochemical papers published by journal.

Journal	No.	Journal	No.
Biochemistry	11,218	ACS Chem. Neurosci.	113
J. Phys. Chem. B	5172	J. Med. Chem.	108
J. Am. Chem. Soc.	4915	Org. Lett.	86
Langmuir	2196	ACS Appl. Mater. Interfaces	54
Biomacromolecules	662	ACS Synth. Biol.	48
J. Chem. Theory Comput.	528	Chem. Mater.	43
J. Phys. Chem. A	528	Mol. Pharmaceutics	42
Inorg. Chem.	512	ACS Cent. Sci.	34
J. Phys. Chem. Lett.	426	J. Nat. Prod.	34
ACS Chem. Biol.	338	J. Chem. Inf. Comput. Sci.	31
Acc. Chem. Res.	320	Environ. Sci. Technol.	22
J. Agric. Food Chem.	282	ACS Biomater. Sci. Eng.	19
Nano Lett.	252	Ind. Eng. Chem. Res.	19
J. Phys. Chem. C	246	ACS Med. Chem. Lett.	18
Chem. Rev.	241	ACS Macro Lett.	15
J. Proteome Res.	233	ACS Sustainable Chem. Eng.	8
Bioconjugate Chem.	231	J. Chem. Eng. Data	8
ACS Nano	205	ACS Catal.	6
J. Chem. Inf. Model.	203	ACS Infect. Dis.	6
Macromolecules	191	ACS Sens.	3
Anal. Chem.	186	ACS Appl. Bio Mater.	2
Chem. Res. Toxicol.	164	ACS Comb. Sci.	2
Crystal Growth and Design	140	J. Chem. Educ.	2
J. Org. Chem.	125	J. Comb. Chem.	2
ACS Omega	122	ACS Earth Space Chem.	1
J. Phys. Chem.	120	Org. Process Res. Dev.	1

Data Preprocessing

The abstracts of the 26,422 papers published in the field of biochemistry were collected and tokenize into units of words. Words that appear after more than 10,000 times, representatively, verbs such as ‘is’, ‘have’, and ‘be’, and unnecessary stopwords including special characters such as punctuation marks, were removed from the data prior to analysis. Then, only the words corresponding to nouns were filtered to be put through the analysis.

5. Results

5.1. Defining the Topics in Biochemical Research

The topic modeling using the preprocessed data proceeded as follows. First, each topic was given a name based on the words assigned to each topic. The number of topics to be analyzed was set to 15, and the outcomes and description of each topic are shown in Table 2. The words assigned to each topic are schematized as Figure 3 using word clouds, and the probability of each topic’s word generation is summarized in Table 3. Figure 4 shows the ratio each topic holds among the total research data.

Table 2. Topic title and descriptions.

#	Title	Descriptions
0	Aberrant protein aggregation and diseases	Amyloid formation; Amyloidosys
1	Ion channels and receptors	Ion channel complexes; Ligands; Drug design
2	Protein folding	Kinetics and thermodynamics of protein folding; Protein denaturation
3	Protein conformation (NMR studies)	NMR spectroscopy for conformational dynamics
4	Various helix dimers	Dimeric transmembrane domain, Dimerizing transcription factors, Zinc-finger domain
5	Protein conformation computational studies)	Molecular dynamics simulations on protein dynamics
6	Membrane transporters	Membrane fusion proteins; Mechanisms of membrane transportation
7	Calculation of electron transfer	Biological electron transfer; Redox pairs
8	Self-assembly of biomolecules	Self-assembled biopolymers; DNA origami; Crystallization
9	Regulation of cellular functions	Signaling pathways; Phosphorylation
10	Biochemistry of lipids	Lipid biochemistry; Vesicle formation; Cholesterols
11	Development of chromophores for biochemistry	Fluorescence sensors; Biological chromophores
12	Biochemistry of nucleic acids	DNA; RNA; DNA secondary structures
13	Biochemistry of Heme complexes	Hemoglobin; Heme complexes
14	Redox chemistry of cytoskeletal dynamics	Kinetics of redox enzymes; Oxidation of actins

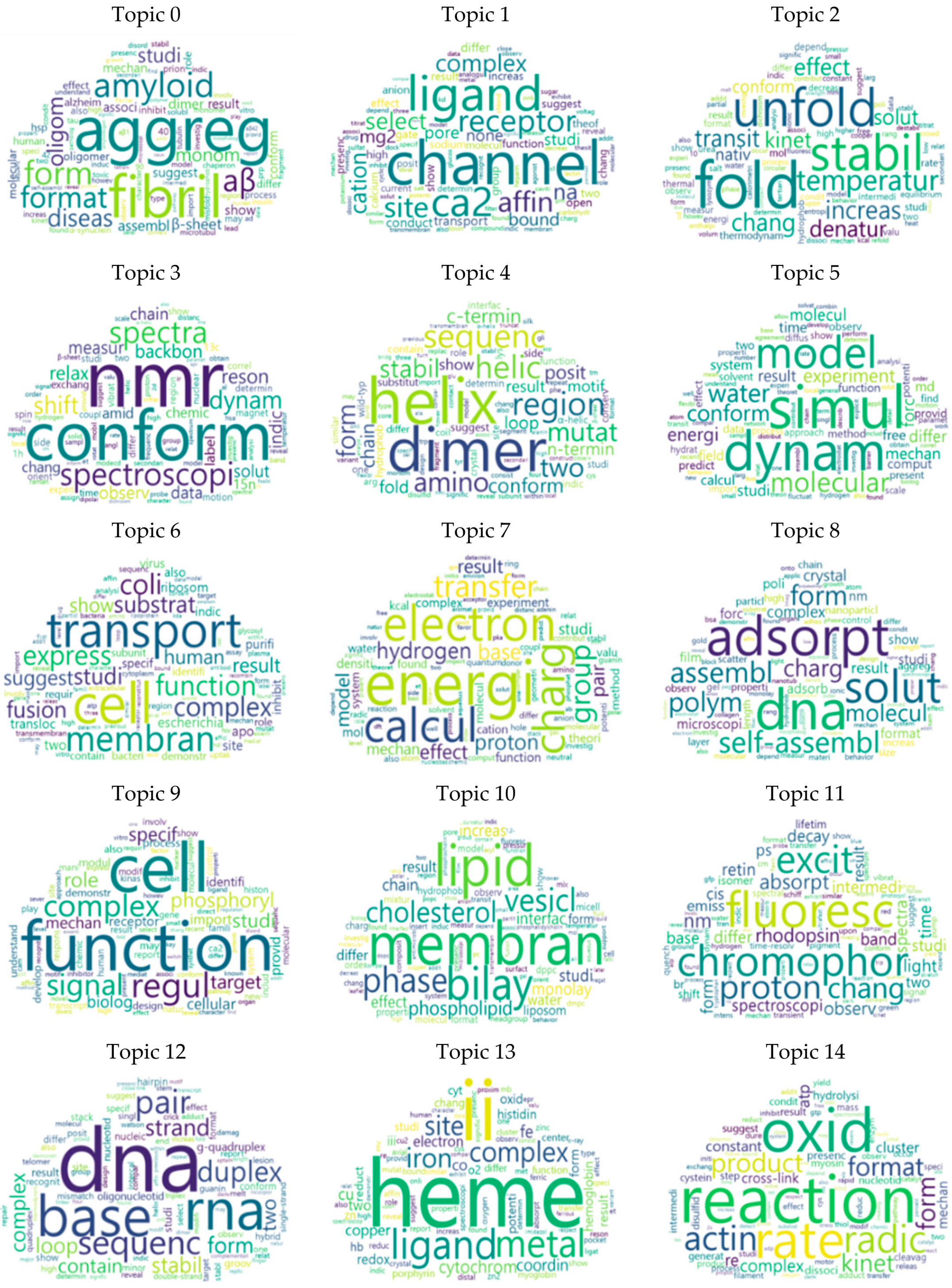

Figure 3. Word cloud by topic.

Table 3. Probability distribution of words by topic.

Aberrant Protein Aggregation and Diseases		Ion Channels and Receptors		Protein Folding		Protein Conformation (NMR Studies)		Various Helix Dimers		Protein Conformation (Computational Studies)		Membrane Transporters		Calculation of Electron Transfer	
aggreg	0.028	channel	0.028	fold	0.026	nmr	0.023	helix	0.012	simul	0.024	cell	0.013	energi	0.018
fibril	0.024	ligand	0.022	unfold	0.016	conform	0.018	dimer	0.011	dynam	0.023	transport	0.013	electron	0.015
amyloid	0.016	ca2	0.015	stabil	0.015	spectro-scopi	0.011	sequenc	0.011	model	0.018	membran	0.011	calcul	0.014
form	0.015	receptor	0.012	temperatur	0.014	spectra	0.011	region	0.010	molecular	0.016	function	0.007	charg	0.013
aβ	0.014	site	0.010	increas	0.011	dynam	0.010	helic	0.010	water	0.013	coli	0.006	transfer	0.012
format	0.014	affin	0.009	chang	0.010	shift	0.009	amino	0.009	conform	0.010	express	0.006	group	0.011
diseas	0.011	complex	0.009	transit	0.010	reson	0.009	mutat	0.009	experiment	0.010	complex	0.006	hydrogen	0.010
oligom	0.009	select	0.009	effect	0.009	relax	0.008	two	0.009	energi	0.010	substrat	0.006	proton	0.009
studi	0.008	cation	0.008	denatur	0.009	backbon	0.008	stabil	0.008	forc	0.008	studi	0.006	base	0.009
monom	0.007	na	0.008	kinet	0.008	measur	0.008	form	0.008	differ	0.007	fusion	0.005	effect	0.008
β-sheet	0.007	mg2	0.007	solut	0.008	indic	0.007	conform	0.008	molecul	0.007	suggest	0.005	pair	0.007
assembl	0.007	bound	0.006	conform	0.007	observ	0.007	c-termin	0.008	mechan	0.007	result	0.005	model	0.007
dimer	0.006	pore	0.006	nativ	0.007	solut	0.007	posit	0.007	result	0.007	human	0.005	result	0.007
show	0.006	none	0.006	thermo-dynam	0.007	chain	0.007	chain	0.007	studi	0.006	show	0.004	mechan	0.006
result	0.005	studi	0.006	studi	0.007	data	0.007	n-termin	0.007	time	0.006	escherichia	0.004	studi	0.006
associ	0.005	transport	0.006	mol	0.007	chang	0.007	loop	0.007	calcul	0.005	two	0.004	complex	0.006
suggest	0.005	conduct	0.005	result	0.006	amid	0.007	fold	0.007	md	0.005	apo	0.004	mol	0.006
mechan	0.005	differ	0.005	rate	0.006	chemic	0.007	result	0.006	system	0.005	transloc	0.004	experiment	0.006
conform	0.005	show	0.005	depend	0.006	15n	0.007	motif	0.005	free	0.005	ribosom	0.004	function	0.005
oligomer	0.005	two	0.005	measur	0.005	label	0.006	show	0.005	field	0.005	site	0.004	densiti	0.005

Self-Assembly of Biomolecules		Regulation of Cellular Functions		Biochemistry of Lipids		Develoment of Chromophores for Biochemistry		Biochemistry of Nucleic Acids		Biochemistry of Heme Complexes		Redox Chemistry of Cytoskeletal Dynamics	
dna	0.009	function	0.013	membran	0.045	fluoresc	0.020	dna	0.064	heme	0.022	reaction	0.022
adsorpt	0.009	cell	0.012	lipid	0.042	chromophor	0.014	rna	0.020	ii	0.017	oxid	0.018
solut	0.008	regul	0.008	bilay	0.025	excit	0.012	base	0.019	ligand	0.014	rate	0.014
self-assembl	0.008	complex	0.008	phase	0.014	proton	0.010	sequenc	0.015	metal	0.014	radic	0.009
assembl	0.008	signal	0.007	vesicl	0.012	chang	0.009	pair	0.014	complex	0.010	actin	0.008
form	0.007	phosphoryl	0.007	cholesterol	0.009	absorpt	0.009	duplex	0.013	iron	0.010	product	0.008
polym	0.007	role	0.006	phospholipid	0.009	nm	0.008	strand	0.008	site	0.010	format	0.008
charg	0.006	studi	0.006	studi	0.007	rhodopsin	0.007	form	0.007	cytochrom	0.010	form	0.007
molecul	0.006	target	0.006	monolay	0.007	observ	0.006	stabil	0.007	fe	0.009	kinet	0.007
complex	0.006	specif	0.006	chain	0.006	light	0.006	loop	0.007	cu	0.009	complex	0.007
crystal	0.006	biolog	0.006	increas	0.006	time	0.006	complex	0.006	coordin	0.009	re	0.006
result	0.006	mechan	0.005	interfac	0.006	retin	0.006	contain	0.006	co	0.008	atp	0.006
microscopi	0.006	cellular	0.005	effect	0.006	differ	0.006	two	0.006	copper	0.007	constant	0.006
studi	0.005	import	0.005	result	0.006	intermedi	0.005	g-quadruplex	0.005	two	0.007	cluster	0.006
forc	0.005	provid	0.005	water	0.005	band	0.005	result	0.005	electron	0.007	cross-link	0.006
format	0.005	receptor	0.005	differ	0.005	spectroscopi	0.005	studi	0.005	redox	0.006	mechan	0.006
adsorb	0.005	identifi	0.004	liposom	0.004	form	0.005	oligonucleotid	0.005	iii	0.005	hydrolysi	0.005
poli	0.005	process	0.004	form	0.004	decay	0.005	site	0.005	form	0.005	presenc	0.005
nm	0.005	modul	0.004	hydrophob	0.004	spectra	0.005	groov	0.005	histidin	0.005	dissoci	0.005
film	0.005	demonstr	0.004	show	0.004	base	0.005	nucleotid	0.005	oxid	0.005	result	0.005

Topic modeling intuitively assigns research fields and topics based on the composition of the words that are assigned to the topic. For example, the words assigned to the first topic are "aggreg, fibril, amyloid, format, diseas", based on which it becomes possible to induce the research topic, "Aberrant protein aggregation and diseases". The second topic, "Ion channels and receptors", was induced based on the words "channel, ligand, receptor, affin, complex" which were assigned to the topic.

Although the topics showed little difference in their ratios among the total research data, the topic that accounted for the highest ratio were "5. Protein conformation (computational studies), 9. Regulation of cellular functions, 10. Biochemistry of lipids", while those which took up comparatively lower portions were "3. Protein folding, 11. Development of chromophores for biochemistry, 14. Redox chemistry of cytoskeletal dynamics".

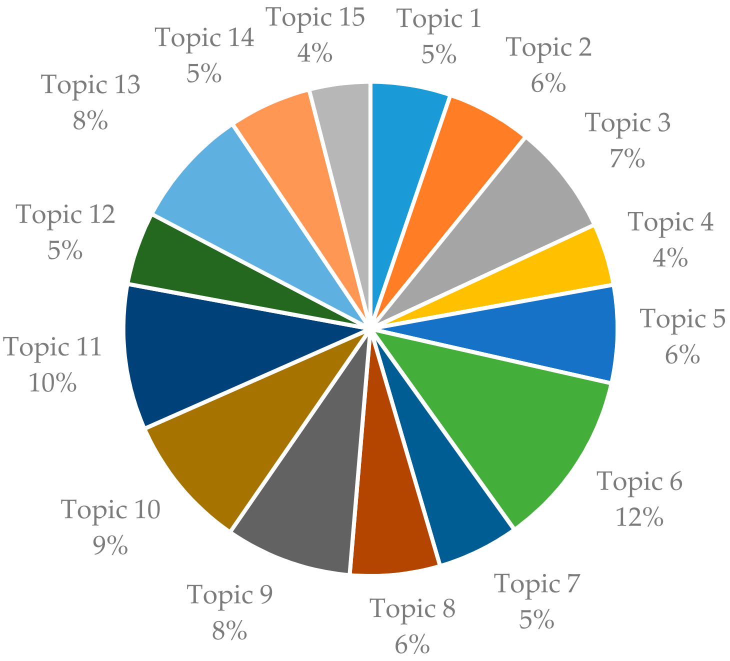

Figure 4. Ratio by topic.

5.2. Analysis of Yearly Trends

Using the topic modeling analysis results for the research conducted in the field of biochemistry from 1999 to 2018, the changes in research trends over time were identified. As research in 2019 is ongoing, this year was excluded from the trend analysis. The overall trend was determined by plotting the ratio of research papers by topic for all years on a graph (Figure 5), and then the data was analyzed quantitatively using the linear regression model (Table 4).

Sun and Yin [13] collected the data on transportation research from 1991 to 2015 for topic modeling analysis and, to analyze the trends in research, defined the r_k index using the ratio of topic k by journal, θ_k^t, following the equation below. Based on the equation, topics whose r_k value is less than 1 were classified as hot topics, and those above 1 as cold topics.

$$r_k = \frac{\sum_{t=1991}^{1995} \theta_k^t}{\sum_{t=2011}^{2015} \theta_k^t} \tag{4}$$

However, this method of analysis is limited in reflecting the overall trends as it is based on simple arithmetic averages, such as r_k. As such, in this study, a linear regression model was constructed by the method proposed by Griffiths and Steyvers [15], and hot and cold topics were classified based on the significance of the regression coefficient. The independent variables were set as the 20 years from

1999 to 2018, and the dependent variables as the share of each topic by year. Topics with regression coefficients under them were considered significant, and whether they were hot or cold topics were judged depending on the direction of the regression coefficient, that is, if (+) was assumed as indicating a hot topic, and (−) as indicating a cold topic.

According to the trend analysis based on linear regression, "0. Aberrant protein aggregation and diseases, 5. Protein conformation (computational studies), 9. Regulation of cellular functions, 10. Biochemistry of lipids" were the topics that showed consisted upward trends. "3. Protein conformation (NMR studies), 4. Various helix dimers, 12. Biochemistry of nucleic acids, 13. Biochemistry of Heme complexes, 14. Redox chemistry of cytoskeletal dynamics" were the topics that exhibited clear downward trends over time throughout the whole period examined in this research.

These trends indicate that biochemical research has gradually broadened its scope from its past focus on the understanding of specific biomolecules (proteins, nucleic acids, etc.) to examining all proteins or lipids, then to explaining globally-occurring phenomena. Interestingly though, the research on lipids has become even more active, which may be a reflection of the recent spotlight on autophagy and lipid droplets leading to a growth in research. Also, the decline in research on the biochemistry of nucleic acids despite the large number of papers published on genome editing technology can be attributed to the research focus on the use of genome editing to control cell or protein functions, rather than nucleic acids themselves. Also notable is that in the case of protein conformational dynamics, computational study (topic 5) is a hot topic, while NMR study (topic 3) is a cold topic. One reason for this can be found in the developments in computational power, both in terms of hardware and software, which has enabled analysis of areas that were out of reach using nuclear magnetic resonance (NMR) technology, leading to a growth trend.

Analysis of Topics by Period

Topic modeling creates word clusters from a given document data set based on probability, then attributes topics to each cluster afterwards. Thus, naturally, a change in the data set results in a change in the word clusters. The topics listed in Table 2 are based on the full data set collected for this study for the twenty year period from 1999 to 2018, which means that dividing this data set by period and applying topic modeling to each periodic set individually will extract topics that are specific to the period, as opposed to the general topics found for the twenty year timeframe. To do this, we divided the data set into four specific timeframes and performed topic modeling separately on each of the four data sets (Table 5). The four timeframes were from 1999 to 2006, 2007 to 2011, 2012 to 2017, and the one-year timeframe of 2018 to identify the latest topics in biochemical research. The number of topics was set to five.

The results showed that regardless of the period, a large number of research has been consistently conducted on topics such as "membran, lipid" and "DNA." On the other hand, "heme" appeared as one of the major topics before 2011 but was not seen thereafter, while "fibril," and "dynam, simul" emerged as a popular topic from 2007. All of these topics are among the fifteen major topics extracted for the full twenty-year period, however, their inclusion in the major topics by period differed by topic and period in a way that matched the trends shown in Figure 5. Also, the topics extracted for 2018 confirmed the appearance of "water, hydrogen, charg" as new topics. The emergence of "energy production in biosystems" as a new topic around 2018 can be tied to the wider implementation of policies in many countries toward accelerating the development of biomass energy.

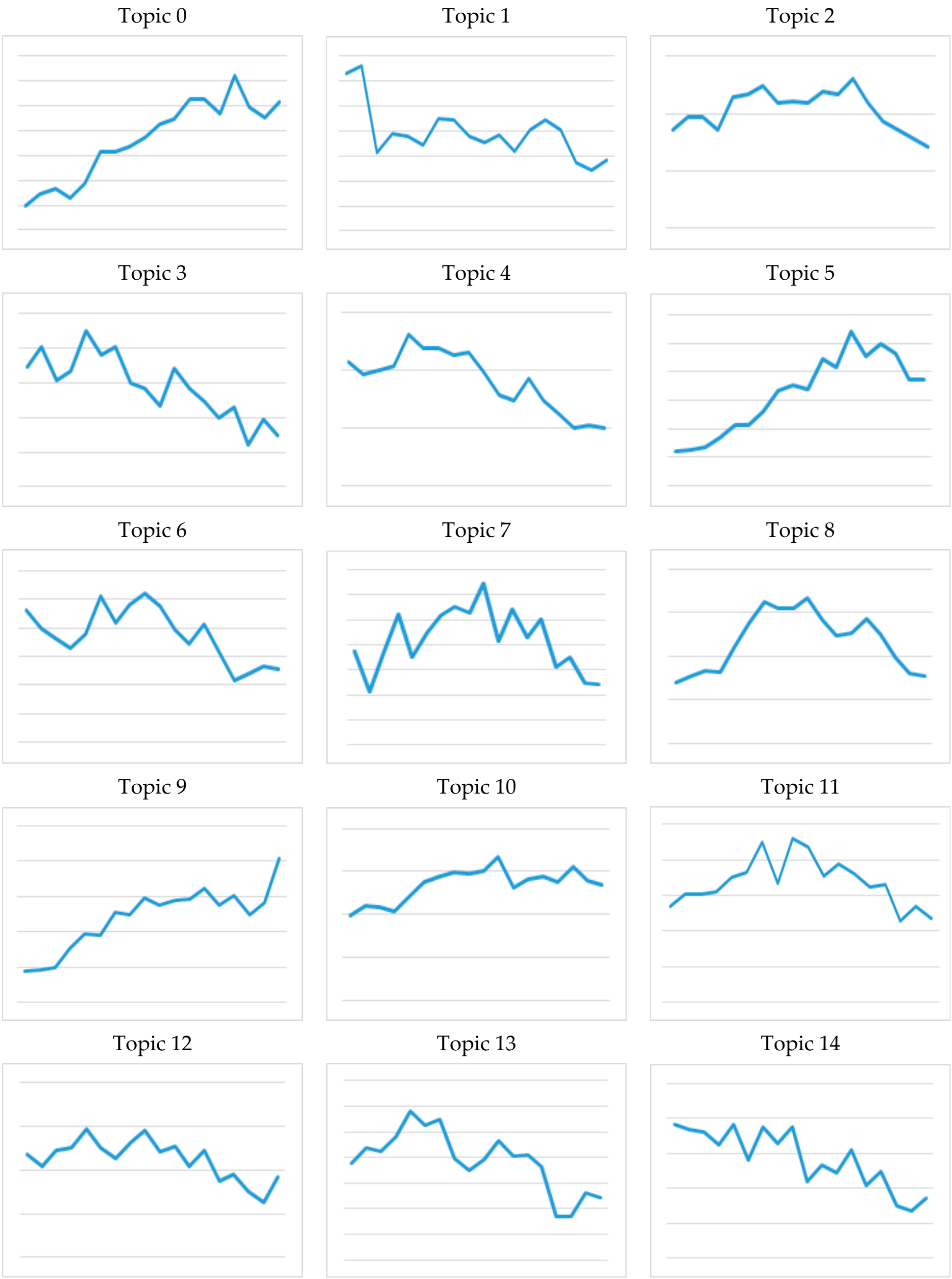

Figure 5. Trends in research topics by year.

Table 4. Hot/Cold topics.

Topic No.	Coefficient	*p*-Value	Hot/Cold
0	5.525	0.0000	hot
1	−2.538	0.0055	-
2	−0.552	0.4843	-
3	−2.838	0.0000	cold
4	−3.934	0.0000	cold
5	11.023	0.0000	hot
6	−2.133	0.0072	-
7	−0.672	0.5599	-
8	0.860	0.5855	-
9	7.128	0.0000	hot
10	2.361	0.0029	hot
11	−0.495	0.4312	-
12	−2.733	0.0041	cold
13	−3.083	0.0012	cold
14	−2.704	0.0000	cold

Table 5. Topics and Words by Timeframe.

Year	Title	Words
1999~2006	Biochemistry of Heme complexes	heme, rate, oxid, reaction, electron
	Biochemistry of lipids	lipid, membran, bilay, phase, water
	Biochemistry of nucleic acids	dna, base, structure, energi, complex
	Membrane transporters	cell, complex, site, function, receptor
	Protein conformation (NMR studies)	structur, fold, conform, nmr, helix
2007~2011	Membrane transporters	cell, function, complex, receptor
	Aberrant protein aggregation and diseases	form, fibril, aggreg, complex, heme
	Biochemistry of nucleic acids	dna, dynam, conform, simul, fold
	Biochemistry of lipids	membran, lipid, bilay, water, phase
	Calculation of electron transfer	base, proton, eletron, energi, transfer
2012~2017	Calculation of electron transfer	eletron, proton, reaction, transfer, complex
	Biochemistry of nucleic acids	dna, rna, base, sequenc, complex
	Aberrant protein aggregation and diseases	aggreg, fibril, conform, cell, amyloid
	Biochemistry of lipids	membran, lipid, bilay, phase
	Protein conformation (computational studies)	dynam, simul, conform, water, fold
2018	Ion channels and receptors	cell, function, receptor, rna, ligand
	Biochemistry of lipids	membran, lipid, bilay, phase, system
	Biochemistry of nucleic acids	dynam, conform, dna, fold, simul
	Energy production in biosystems	water, energi, hydrogen, chain, charg
	Aberrant protein aggregation and diseases	aggreg, form, fibril, cell, diseas

6. Discussion and Conclusions

Since its establishment as a separate field of study, biochemistry has been continuously expanding in its research scope, and the related industries have also shown steady growth which is expected to continue in the future [10]. Today, the biochemical industry continues to be of great attraction not only to existing biotech and chemical firms but also inviting new challengers including global companies such as Coca-Cola, IKEA, Dell, and LEGO, who are also preparing to try their hands in the research and development of biochemical products. Meanwhile, countries have also been actively supporting the research and development of biochemistry through national policies. The United States (US) plans to replace 30% of its current oil consumption with green carbon by 2030 and is supporting the wider use of bio-derived products policy-wise by expanding its Biopreferred Program to 97 items and 10,000 types. The European Union, which accounts for 60% of the global bioplastics market, has been developing the biochemical industry as one of its six leading industries, and in Japan, national efforts are being made to promote the biochemical industry through the country's "Biomass Japan

Comprehensive Strategy" [30]. This close connection between research and industry has been a driver for biochemistry to develop at greater speed as well as embrace new areas and topics, which is why the present study's application of topic modeling, which is often attempted in other studies in diverse fields, for the analysis of biochemical research trends can contribute to existing literature [2–5].

This study used topic modeling, a text mining technique, based on LDA to define the research topics in biochemical research over the past twenty years and quantitatively analyze their trends. The abstracts of 26,422 papers published in 52 journals from 1999 to 2018 were collected through the American Chemical Society and used as the data for analysis. Based on the results, we identified the fifteen major topics of biochemistry over the past 20 years and, using linear regression, analyzed the amount of research conducted over time. Further, the research data was divided into four periods to repeat the topic modeling analysis for the specific timeframes to see which topics decreased or increased in weight over time and to pinpoint newly-emerging topics.

Our analysis results were in line with the recent trends of the biochemical industry. As recent movements in the industry—such as the 1300-PDO production plant with a capacity of 450,000 tons per year constructed by Dupont and Tate and Lyle to produce PTT(Polytrimethylene terephthalate) to be used as fiber material for the Sorona brand and the joint venture for Propylene glycol (PG) production established by ADM(Archer Deniels Miland) and Cargill—the attention of the industry is pointed at biofiber and biofuels. The latest trends revealed by our study lists "energy production in biosystems" and "aberrant protein aggregation and diseases" as two of the top five topics, indicating that our analysis closely reflects the latest industrial interests. Therefore, applying the analysis method used in this study with continuously updated data will provide a helpful decision-making tool for practitioners and researchers in the industry and academia.

Most researchers, of course, do not seek to follow the trend of the moment, nor should they. The goal of this study is not to argue that researchers should follow academic trends but to contribute to future research by sharing information on trends and opening up possibilities of new and diverse research. It is important to study classic topics to gain an understanding of the fundamentals of a field, but it is also equally necessary to bring together different research fields or to broaden boundaries and uncover new topics according to changing trends or technological development, thereby breathing in new vitality to traditional research topics.

In addition to the papers published by the ACS in this study, there exists a number of journals, including Nature, which publish articles that can be categorized under the field of biochemistry. However, we used only the papers clearly categorized as biochemistry by the ACS to avoid the subjective judgment of what research belongs to the field. Our decision here is both a limitation and a contribution of this study. We put aside research articles which were unspecified as to whether they fall under biochemistry, however, this also means the topics extracted by our analysis can become a base-point from which further trend analyses can be performed with additional data. In particular, further studies may expand our present study by conducting a more in-depth analysis of sub-categories to uncover more specific trends in the wide scope of biochemical research.

Author Contributions: C.K. collected the data and H.J.K. conceived and designed the experiments. K.K. provided biochemical knowledge. The experiment was performed by all related authors. Also, the paper is written by all related authors. All authors read and approved the final manuscript.

Funding: This work was funded by Incheon National University "Program on Research Workforce Post-Doctor Recruitment Support Project 2019".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. American Chemical Society. Available online: www.Acs.org (accessed on 8 April 2019).
2. McKendry, P. Energy production from biomass (part 1): Overview of biomass. *Bioresour. Technol.* **2002**, *83*, 37–46. [CrossRef]

3. McKendry, P. Energy production from biomass (part 2): Conversion technologies. *Bioresour. Technol.* **2002**, *83*, 47–54. [CrossRef]
4. McKendry, P. Energy production from biomass (part 3): Gasification technologies. *Bioresour. Technol.* **2002**, *83*, 55–63. [CrossRef]
5. Galambos, L.; Takashi, H.; Vera, Z. (Eds.) *The Global Chemical Industry in the Age of the Petrochemical Revolution*; Cambridge University Press: Cambridge, UK, 2007.
6. Bull, A.T.; Holt, G.; Malcolm, D.L. *Biotechnology: International Trends and Perspectives*; OCDE: Paris, France, 1982.
7. Demirbas, A. Combustion characteristics of different biomass fuels. *Prog. Energy Combust. Sci.* **2004**, *30*, 219–230. [CrossRef]
8. Fetterhoff, T.J.; Voelkel, D. Managing open innovation in biotechnology. *Res. Technol. Manag.* **2006**, *49*, 14–18. [CrossRef]
9. Kinch, M.S. The rise (and decline?) of biotechnology. *Drug Discov. Today* **2014**, *19*, 1686–1690. [CrossRef] [PubMed]
10. Jian, Z.; Zhao, Z.-Y. Green building research–current status and future agenda: A review. *Renew. Sustain. Energy Rev.* **2014**, *30*, 271–281.
11. William, L.B.; Mickelsen, J.F. An analysis of prior Delphi applications and some observations on its future applicability. *Technol. Forecast. Soc. Chang.* **1977**, *10*, 103–110.
12. Clark, K.R.; Neal, T.A.; Johnson, T.E. Creation of an innovative laser incident reporting form for improved trend analysis using the Delphi technique. *Mil. Med.* **2006**, *171*, 894–899. [CrossRef] [PubMed]
13. Sun, L.; Yin, Y. Discovering themes and trends in transportation research using topic modeling. *Transp. Res. Part C Emerg. Technol.* **2017**, *77*, 49–66. [CrossRef]
14. Steyvers, M.; Griffiths, T. Probabilistic topic models. *Handb. Latent Semant. Anal.* **2007**, *427*, 424–440.
15. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef] [PubMed]
16. Wang, C.; Blei, D.; Heckerman, D. Continuous time dynamic topic models. *arXiv* **2012**, arXiv:1206.3298.
17. Grimmer, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Anal.* **2010**, *18*, 1–35. [CrossRef]
18. Mann, G.S.; Mimno, D.; McCallum, A. Bibliometric impact measures leveraging topic analysis. In Proceedings of the 6th ACM/IEEE-Cs Joint Conference on Digital libraries, Chapel Hill, NC, USA, 11–15 June 2006.
19. Newman, D.J.; Block, S. Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 753–767. [CrossRef]
20. Gerrish, S.; Blei, D.M. A Language-based Approach to Measuring Scholarly Impact. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
21. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
22. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
23. Foltz, P.W.; Kintsch, W.; Landauer, T.K. The measurement of textual coherence with latent semantic analysis. *Discourse Process.* **1998**, *25*, 285–307. [CrossRef]
24. Landauer, T.K.; Dumais, S.T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **1997**, *104*, 211. [CrossRef]
25. Landauer, T.K.; Foltz, P.W.; Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **1998**, *25*, 259–284. [CrossRef]
26. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608. [CrossRef] [PubMed]
27. Lee, S.; Song, J.; Kim, Y. An empirical comparison of four text mining methods. *J. Comput. Inf. Syst.* **2010**, *51*, 1–10.
28. Bergamaschi, S.; Po, L. Comparing lda and lsa topic models for content-based movie recommendation systems. In Proceedings of the International Conference on Web Information Systems and Technologies, Barcelona, Spain, 3–5 April 2014.

29. Cvitanic, T.; Lee, B.; Song, H.I.; Fu, K.; Rosen, D. Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents. In Proceedings of the International Conference on Case-Based Reasoning, Atlanta, GA, USA, 31 October–2 November 2016.

30. Lee, M. Bio-Based Chemicals Industry Trends. In *Bio Economy Brief*; Korea Biotechnology Industry Organization: Seoul, Korea, 2018; pp. 2–3.