



Identification of Intra-Domain Ambiguity using Transformer-based Machine Learning

Ambarish Moharil
Dept. of Instrumentation & Control Engineering
Vishwakarma Institute of Technology
Pune, India
Email: ambarish.m02@gmail.com

Arpit Sharma
Dept. of Electrical Engineering & Computer Science
Indian Institute of Science Education and Research
Bhopal, India
Email: arpit@iiserb.ac.in

Experimentation Journal for the C'_{cs} Corpus (Computer Science Corpus)



Table of Contents

| Content | Page Number |
|--|--------------------|
| Section 1 :- Experiments from the paper | Pg 3-25 |
| Application | Pg 3-7 |
| Function | Pg 7-11 |
| Language | Pg 11-15 |
| Cell | Pg 15-20 |
| Object | Pg 20-25 |
| Section 2 :- Some more words | Pg 25-53 |
| Operation | Pg 25-30 |
| Source | Pg 30-35 |
| State | Pg 35-39 |
| Table | Pg 39-43 |
| Window | Pg 43-46 |
| Machine | Pg 46-53 |
| <u>Section 3:- Overlapping Clusters</u> | Pg 53-61 |
| Interface | Pg 53-57 |
| System | Pg 57-61 |



Computer Science Domain Result Analysis

Section 1:- Experiments From The Paper

This report consists of the results obtained after experimentations on the C'_{CS} corpus. We divide this report in three sections. Section 1 explains the analysis for the words selected in the paper. Section 2 lists a few more examples and Section 3 consists of the examples of the words which form overlapping clusters.

The analysis for each word begins with the study of the “*Elbow Graph*” followed by a semantic contextual analysis of every cluster.

Each cluster consists of a threshold plot at the end, which portrays the cosine similarity between the selected target word and all the other words in that respective cluster. In order to select the relevant context terms associated with that particular target label, we define a threshold which is experimentally computed as 0.45 (quite visible from the various threshold plots).

The final results are mentioned in Section V of our paper and the reader can refer to a profound analysis of the same in this report.

Target word: - Application

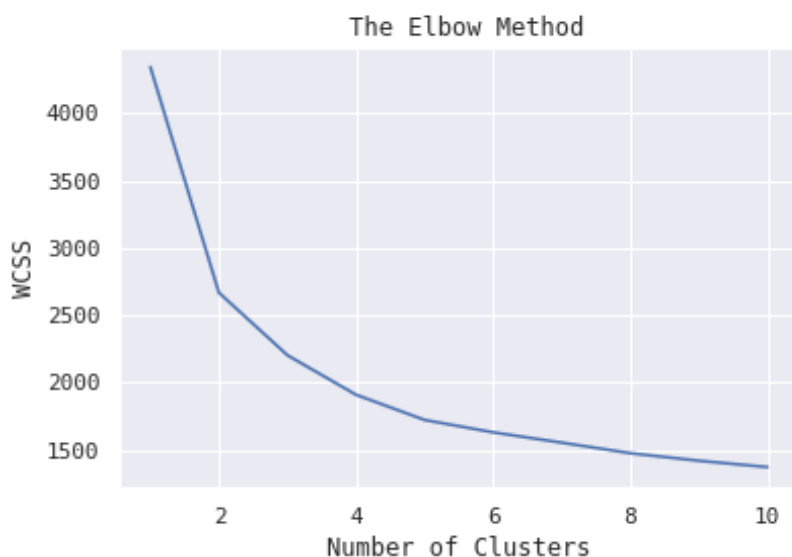
No of clusters: - 2

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Method

Cluster 0

Implied Meaning in the Cluster: - “App”, “Program”, “Software”.

The word application has been divided into two clusters by the k-means algorithm.

This shows that the word has been used in 2 contexts in the corpus that was used for the analysis.

The clusters seem quite discernable altogether. The first cluster implies the use of the word application as a “**software application**” or an “**app**”.

In this cluster the word “**application**” **is** used in noun phrases like:-

“Payroll Application”, “Application Software”, “Database Application Interface”, “Web Applications”, “Application Development”, “Android Applications”, “Windows Applications”, “Computer applications”, “Ping application”.

Example Sentences: -

1) Database **payroll application** also initiates the process of exchanging the data of a ship shore, the data transfer task often contains many components. These layers produce the effect of isolating the implementation details of on one component assorted with the internal methods others.

2) Thin clients use resources host computer. A thin client generally presents processed data provided by the application server, and performs bulk required data processing. A device using **web application** such as Office Web Apps is a thin client.

3) System Development Methodology refers to a standard process followed in an organization to conduct all the steps necessary to analyze, design, implement, and maintain information systems. [1] There are a number of methodologies used to develop and improve the system such as the Traditional Waterfall, **Rapid Application Development**, Prototyping, Agile and etc. However, this paper focuses only on the agile methodologies and their specific techniques, people's perspectives and the best practice.

4) Ping is a computer network administration **application** used to test the reachability of a host on an Internet Protocol (IP) network. It is available for virtually all operating systems that have networking capability, including most embedded network administration software.

Context Words: -

software, app, program, interface, users, deployment, rest, bot, tool, website etc.

Threshold Analysis: -

Comprehension Window

(No of relevant context words required to comprehend the context of the target word) :- 6

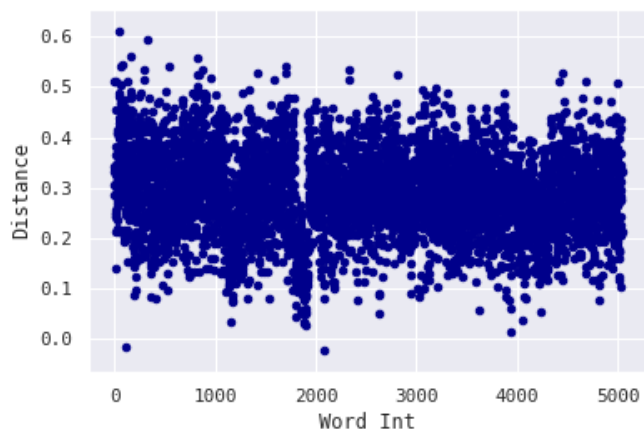
Max Difference

(Difference from the closest and the farthest relevant word. This provides the range of relevant words): - 0.23

Distance of the farthest relevant word : - 0.48

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-----------------|--------|--------|--------|
| Application 425 | 0.286 | 0.626 | 0.2870 |
| Application 420 | 0.320 | 0.6085 | 0.321 |
| Application 106 | 0.2988 | 0.540 | 0.3012 |
| Application 69 | 0.3561 | 0.6448 | 0.366 |



Plot for label 4 and label 227 for cluster 0

Cluster 1

Implied meaning in the cluster :- “Implementation”, “Use”, “Apply”, “Usability”.

In this cluster the word **application** is used in the context of applying or in the sense of **utility and application**. The word can be found used in the noun phrases like “ demonstrates application”, “practical applications of algorithms”, “application of computational tools”, “ modelling applications”, “bioinformatics applications”.

Example Sentences: -

- 1) Computational anatomy is a discipline focusing on the study of anatomical shape and form at the visible or gross anatomical $50 - 100 \mu$ scale of morphology. It involves the development and **application of** computational, mathematical and data-analytical methods for modeling and simulation of biological structures. It focuses on the anatomical structures being imaged, rather than the medical imaging devices.
- 2) A computer scientist specializes theory of computation design and computational systems. [29] Its subfields can be divided into practical techniques for its implementation and **application** in computer systems and purely theoretical areas
- 3) **Depending on the applications application**, one-way delay round-trip time has a significant impact on the performance. Depending on the installation requirements, network performance is usually measured by quality service of telecommunications products. The parameters that affect typically include throughput, jitter, bit error rate latency.
- 4) Recently, **application** of neuroscience research results has also given rise to applied disciplines like neuro-economics, neuro-education, neuro-ethics, neuro-law.

Context words: -

applying, applied, implementations, ability, experimentation, optimization, purpose, acceptance, traction, apply, applicable etc.

Threshold Analysis: -

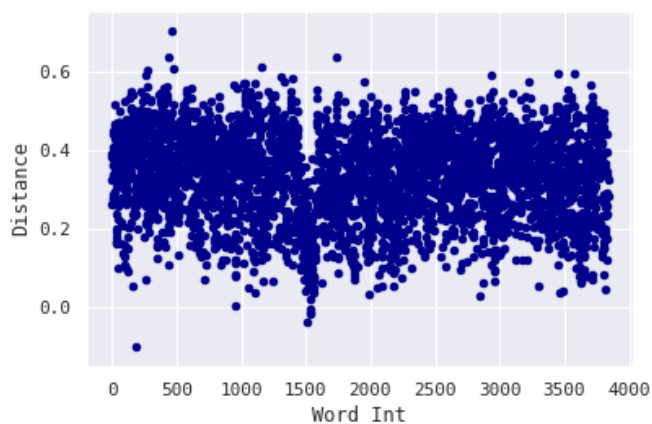
Comprehension Window :- 11

Max Difference : - 0.212

Distance of the farthest relevant word : - 0.507

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-----------------|-------|--------|--------|
| Application 326 | 0.28 | 0.572 | 0.2915 |
| Application 373 | 0.305 | 0.548 | 0.313 |
| Application 44 | 0.33 | 0.702 | 0.348 |
| Application 43 | 0.360 | 0.6135 | 0.3475 |



Plot for label 44 of Cluster 1

Target word: - Function

No of clusters: - 2

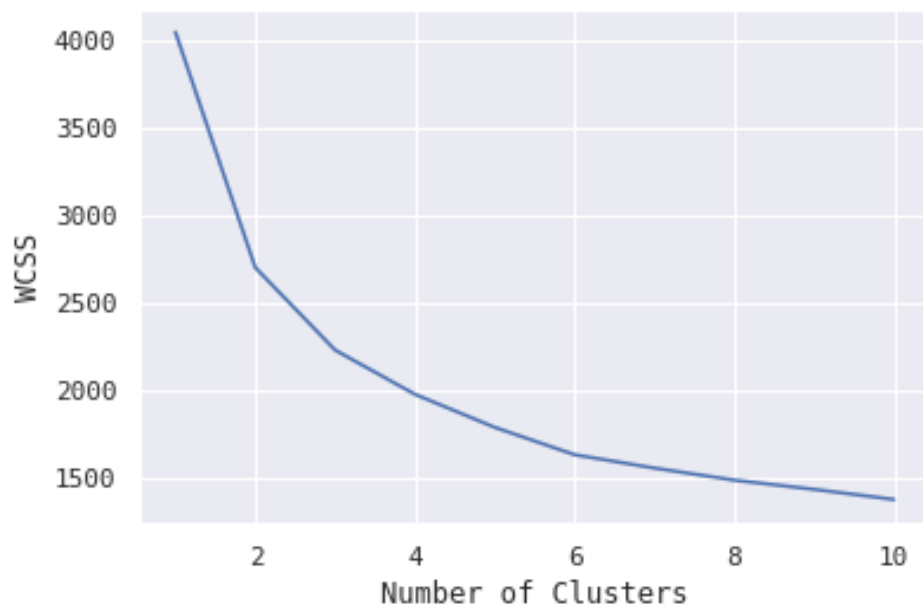
Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES

The Elbow Method



Elbow Method

Correlation Matrix

Cluster 0

Implied Meaning :- “Variable”, “Programming Function”, “Mathematical Function”

In this cluster, the target word *function* is used in the context of a variable, method or a mathematical function. The frequency of the use of the word can be found in sentences that have been used in context of a variable explanation, function call, or describing a mathematical problem or a function.

Example Sentences: -

- 1) Another way to define an ADT, closer to the spirit of functional programming, is to consider each state of the structure as a separate entity. In this view, any operation that modifies the ADT is modeled as a **mathematical function** that takes the old state as an argument and returns the new state as part of the result.
- 2) While a graphical interface for an email client might provide a user with a button that performs all the steps for fetching and highlighting new emails, an API for file input/output might give the developer **a function** that copies a file from one location to another without requiring that the developer understand the file system operations occurring behind the scenes.
- 3) This function sums three terms: $6x^4$, $-2x^3$, 5 Of these three terms one has the highest growth rate and one the largest exponent function x , namely $6x^4$.

4). Continuous probability distribution, See also: **Probability density function**. A continuous probability distribution is a cumulative distribution function which is absolutely continuous.

Context Words:-

variable, curve, vectors, program, mathematical, loop, call, method, class, polynomial, equations, tables, structure etc.

Threshold Analysis: -

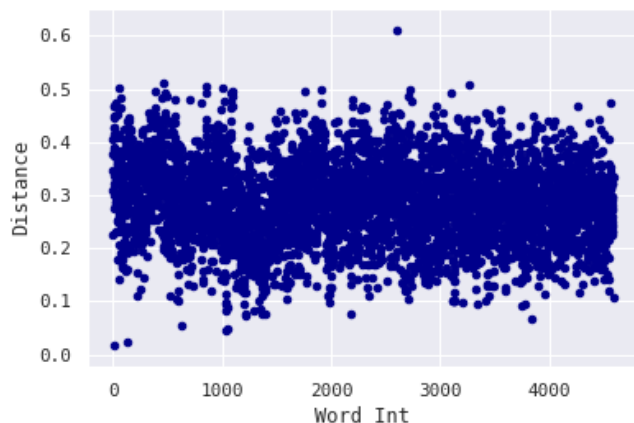
Comprehension Window :- 6

Max Difference : - 0.1678

Distance of the farthest relevant word : - 0.4599

Analysis of word vectors

| Word Label | Mean | Max | Median |
|--------------|--------|--------|--------|
| Function 247 | 0.288 | 0.557 | 0.2695 |
| Function 234 | 0.2868 | 0.6115 | 0.2876 |
| Function 819 | 0.2975 | 0.5659 | 0.3083 |
| Function 236 | 0.2951 | 0.6017 | 0.2992 |



Plot for label 235 of cluster 0

Cluster 1

Implied Meaning: - "utility", "operation", "execution", "procedure"

In this cluster the word function has been used in the context of utility, procedure or a heuristic. A particular operation or a method to accomplish a particular task.



Example Sentences: -

1): Physiology is a study of mechanical, physical, biochemical processes make living organisms **function** as a whole. Physiological studies are traditionally divided into plant physiology animal physiology, principles physiology universal, matter particular organism studied

2) Many copy protection schemes include software blacklisting. The company Password RBL offers password blacklist Microsoft's Active Directory, web sites apps, distributed via RESTful API. Members online auction sites may add members personal blacklist. This means cannot bid ask questions auctions, use "buy now" **function** items.

3) Stable synapses forget less easily, also harder consolidate. One recent computational hypothesis involves cascades plasticity allow synapses **function** multiple time scales.

4) It likely computational tools contribute greatly understanding synapses **function change** relation external stimulus coming decades. Behaviors networks [edit] Biological neurons connected complex, recurrent fashion. These connections are, unlike artificial neural networks, sparse usually specific.

Context Words: -

Functioning, interaction, methodology, heuristic, organization, operation, behaviour, interaction, occurrence etc.

Threshold Analysis: -

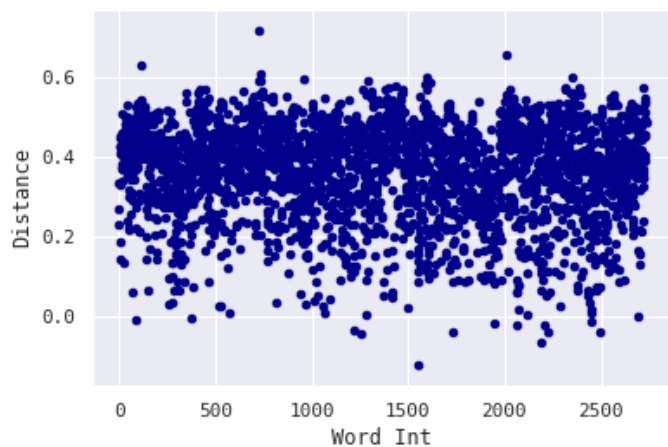
Comprehension Window :- 7

Max Difference : - 0.3601

Distance of the farthest relevant word : - 0.456

Analysis of word vectors

| Word Label | Mean | Max | Median |
|--------------|---------|--------|--------|
| Function 180 | 0.36 | 0.7165 | 0.37 |
| Function 836 | 0.323 | 0.595 | 0.336 |
| Function 255 | 0.29862 | 0.567 | 0.299 |
| Function 357 | 0.370 | 0.690 | 0.38 |



Plot for label 180 and 255 of cluster 1

Target word: - language

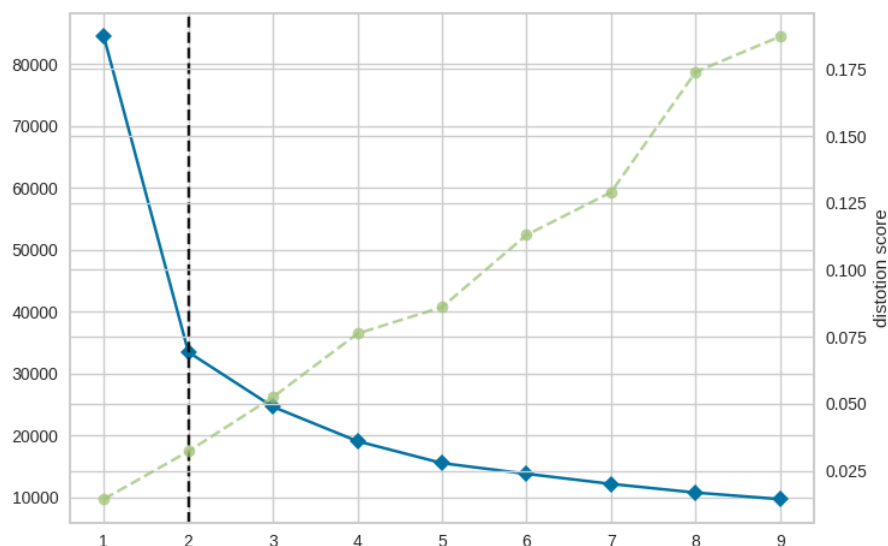
No of clusters: - 2

Deterministic Method :- C-Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Cluster 0

Implied Meaning : - “Programming Language”

Example Sentences:-

1) When analyzing efficiency algorithms use stacks, one may also specify operations take time matter many data items pushed stack, stack uses constant amount storage element. Introduction[edit]Abstract data types purely theoretical entities, used among things simplify description abstract algorithms, classify evaluate data structures, formally describe type systems **programming languages**. However, ADT may implemented specific data types data structures, many ways many **programming languages**; described formal specification language. ADTs often implemented modules: module's interface declares procedures correspond ADT operations, sometimes comments describe constraints.

2) A broad division may drawn "imperative" "functional" definition styles.Imperative-style definition[edit]In philosophy imperative **programming languages**, abstract data structure conceived entity mutable—meaning may different states different times. Some operations may change state ADT; therefore, order operations evaluated important, operation entities may different effects executed different times—just like instructions computer; commands procedures imperative language. To underscore view, customary say operations executed applied, rather evaluated.

3) April 2013Scoping rules representative languages follow.C[edit]Main article: Linkage softwareIn C, scope traditionally known linkage visibility, particularly variables. C lexically scoped **language** global scope known external linkage, form module scope file scope known internal linkage, local scope within function; within function scopes nested via block scope. However, standard C support nested functions.The lifetime visibility variable determined storage class.

4) The applies particular parts program nested functions used, and, naturally, **programs written language** nested functions available such C language.History[edit]Lexical scoping used imperative language ALGOL 60 picked imperative languages since then.

Context Words: - interpreter, syntax, compiler, compile, tool, function, function, code, program, structure, index, indentation, symbolic, logic etc

Threshold Analysis: -

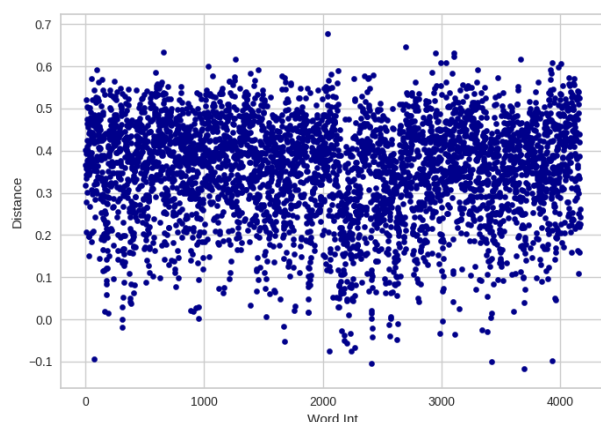
Comprehension Window :- 8

Max Difference : - 0.252

Distance of the farthest relevant word : - 0.489

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Table 224 | 0.365 | 0.689 | 0.365 |
| Table 654 | 0.321 | 0.654 | 0.321 |
| Table 456 | 0.322 | 0.617 | 0.322 |
| Table 214 | 0.362 | 0.559 | 0.3221 |



Threshold scatter plot for the cluster 0.

Cluster 1

Implied Meaning : - “Mode of Communication”

Example Sentences:-

- 1) Automata theory is closely related to **formal language** theory. In this context, automata are used as finite representations of formal languages that may be infinite. Automata are often classified by the class of formal languages they can recognize, as in the Chomsky hierarchy, which describes a nesting relationship between major classes of automata
- 2) To form an idea of the historical place of Jabir's alchemy and to tackle the problem of its sources, it is advisable to compare it with what remains to us of the alchemical literature in the **Greek language**.
- 3) The modern word alchemy in turn is derived from the Arabic word al-kīmīā (الكيمياء). This may have Egyptian origins since al-kīmīā is derived from the Greek χημία, which is in turn derived from the word Kemet, which is the ancient name of Egypt in the **Egyptian language**.
- 4) The theory known as true arithmetic consists of all true statements about the standard integers in the **language of Peano arithmetic**. This theory is consistent, and complete, and contains a sufficient amount of arithmetic. However it does not have a recursively enumerable set of axioms, and thus does not satisfy the hypotheses of the incompleteness theorems.

Context Words: - sentence, grammar, linguistic, sentences, dictionary, context, learned, arguments, grammatical, medium word, words, abstract, clause etc.

Threshold Analysis: -

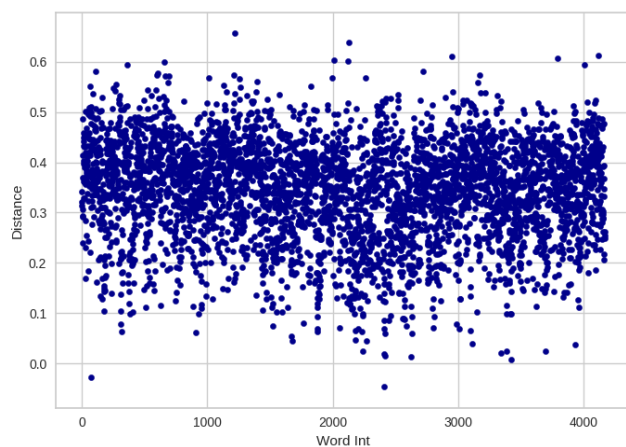
Comprehension Window :- 7

Max Difference :- 0.212

Distance of the farthest relevant word :- 0.475

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Label 254 | 0.315 | 0.654 | 0.32 |
| Label 657 | 0.369 | 0.589 | 0.324 |
| Label 431 | 0.256 | 0.632 | 0.365 |
| Label 270 | 0.227 | 0.599 | 0.321 |



Threshold scatter plot for the cluster

Target word: - Cell

No of clusters: - 3

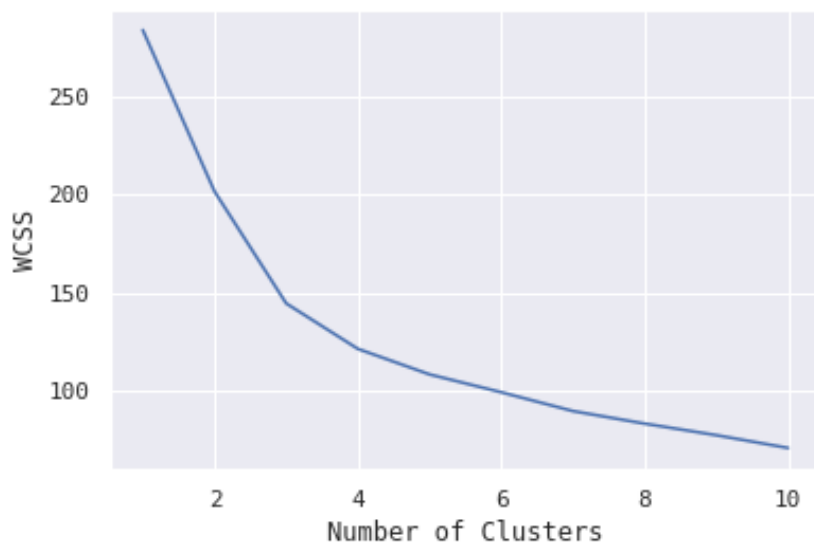
Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES

The Elbow Method



Elbow Method

Cluster 0

Implied Meaning :- “Biological Cell”, “Smallest unit of a human body”.

In this cluster the word cell is used in context of biology. The implications of the word cell are related to biology, *cellular structure, cellular function and cellular specifications, functioning or structure.*

The boundaries and the sentences collected by the algorithm are apt and provide a thorough understanding of the implied context of the target word.

Example Sentences: -

- 1) **Cells** consist of cytoplasm enclosed within a membrane, which contains many biomolecules such as proteins and nucleic acids.
- 2) Advances in microscopy also had a profound impact on biological thinking. In early 19th century, number of biologists pointed the central importance of a **cell**. Then, 1838, Schleiden Schwann began promoting the universal ideas of 1 basic unit of organisms as **cell**.
- 3) Prokaryotes include bacteria and archaea, two of the three domains of life. Prokaryotic cells were the first form of life on Earth, characterized by having vital biological processes including **cell signaling**.
- 4) Enclosing the cell is the cell envelope – generally consisting of a plasma membrane covered by a **cell wall** which, for some bacteria, may be further covered by a third layer called a capsule. Though most prokaryotes have both a cell membrane and a cell wall, there are exceptions such as Mycoplasma (bacteria) and Thermoplasma (archaea) which only possess the cell membrane layer.

Context Words: -

organism, life, root, membrane, chromosomes, cellular, bodies, blood, radiation, molecular, nucleus, tissues, acids, systems, stem etc.

Threshold Analysis: -

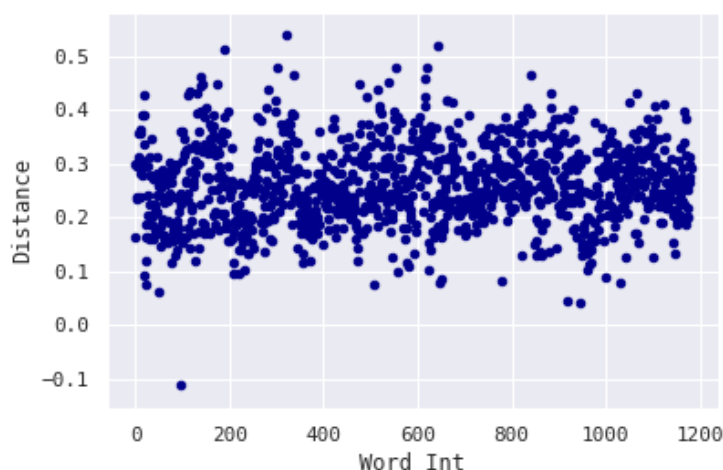
Comprehension Window :- 6

Max Difference : - 0.2535

Distance of the farthest relevant word : - 0.4705

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|--------|--------|
| Cell 13 | 0.2619 | 0.5389 | 0.2577 |
| Cell 15 | 0.2806 | 0.6258 | 0.2779 |
| Cell 28 | 0.318 | 0.6076 | 0.318 |
| Cell 2 | 0.3139 | 0.5933 | 0.3139 |



Plot for label 13 of cluster 0

Cluster 1

Implied Meaning:- “Cell Phone”, “Mobile”, “Portable Communication Device”.

In this cluster the implied meaning of the target word is a portable device or a cell phone. The word is used in the context of functionality, structure, usability and composition of a portable communication device.

Example Sentences: -



1) The wireless revolution began in the 1990s, with the advent of digital wireless networks leading to a social revolution, and a paradigm shift from wired to wireless technology, including the proliferation of commercial wireless technologies such as **cell phones**.

2) Additionally, connected cars may use WiFi Bluetooth to communicate with the onboard consumer devices via **cell phone** network. [43] Self-driving cars expected to be even complex.

3) This includes bullying use email, instant messaging, social media websites such Facebook, text messages, **cell phones**. It stated Cyberbullying common to secondary and primary school.

4) Because of its open license model, the Linux kernel code is available for study and modification, which resulted in its use on a wide range of computing machinery from supercomputers to smart-watches.

Context Words: -

portable, cellular, phone, battery, network, tower, radio, tablet, information, wireless, smart, device, embedded, ring, lan etc.

Threshold Analysis: -

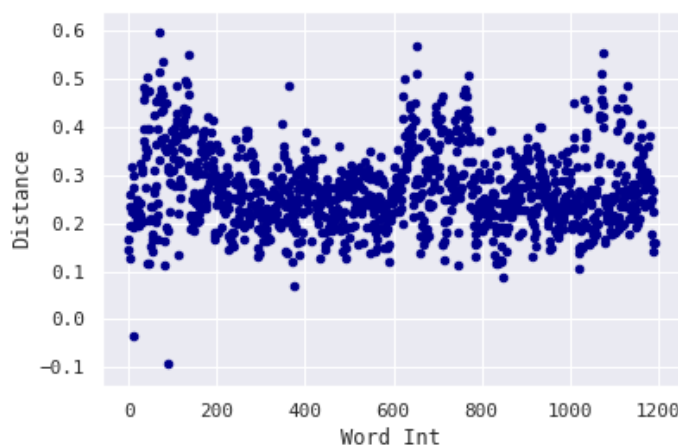
Comprehension Window :-

Max Difference : -

Distance of the farthest relevant word : -

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|---------|---------|
| Cell 94 | 0.315 | 0.621 | 0.31866 |
| Cell 43 | 0.3295 | 0.6704 | 0.3298 |
| Cell 79 | 0.269 | 0.5949 | 0.256 |
| Cell 83 | 0.2829 | 0.60741 | 0.2761 |



Plot for label 79 of cluster 1

Cluster 2

Implied meaning: - “Array”, “Tabular Cell”, “Storage Unit”.

In this cluster the implied meaning of the word cell is with respect to an array. It used in the context of **structuring and documentation**. The sentences and words collected in this clusters are related to tabular representation, statistics etc.

Example Sentences: -

1) For example, it is easy to support COUNT, MAX, MIN, and SUM in OLAP, since these can be computed for **each cell** of the OLAP cube and then rolled up, since on overall sum (or count etc.) is the sum of sub-sums, but it is difficult to support MEDIAN, as that must be computed for every view separately: the median of a set is not the median of medians of subsets.

2) Another very simple implementation technique, usable when the keys are restricted to a narrow range, is direct addressing into an array: the value for a given key k is stored at the array cell $A[k]$, or if there is no mapping for k then the **cell stores** a special **sentinel value** that indicates the absence of a mapping.

3) Modern memory is implemented as semiconductor memory,[3][4] where data is stored within **memory cells** built from MOS transistors on an integrated circuit.

4) Flash memory organization includes both one bit per memory cell and **multi-level cell capable** of storing multiple bits per cell. The memory cells are grouped into words of fixed word length, for example, 1, 2, 4, 8, 16, 32, 64 or 128 bits. Each word can be accessed by a binary address of N bits, making it possible to store 2^N words in the memory.

Context Words: -

results, list, segment, table, matrix, storage, placed, junction, vectors, values center, entity, unit etc.

Threshold Analysis: -

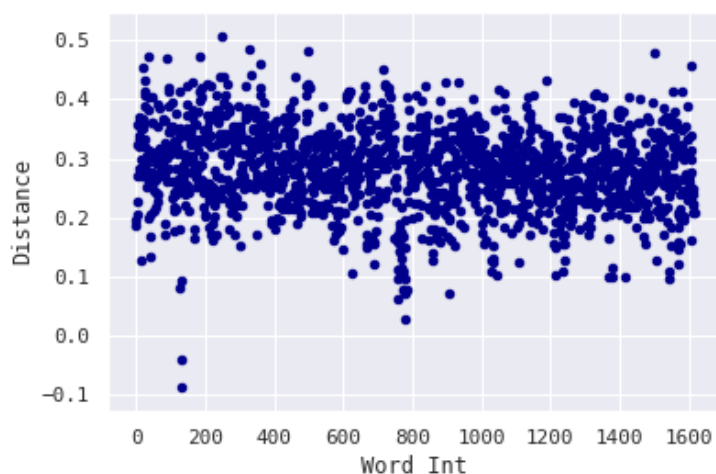
Comprehension Window :- 7

Max Difference : - 0.1823

Distance of the farthest relevant word : - 0.454

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|--------|--------|
| Cell 45 | 0.284 | 0.507 | 0.507 |
| Cell 53 | 0.3506 | 0.6360 | 0.3581 |
| Cell 64 | 0.3369 | 0.5904 | 0.3434 |
| Cell 70 | 0.336 | 0.5916 | 0.3456 |



Plot for label 45 of cluster 2

Target word: - Object

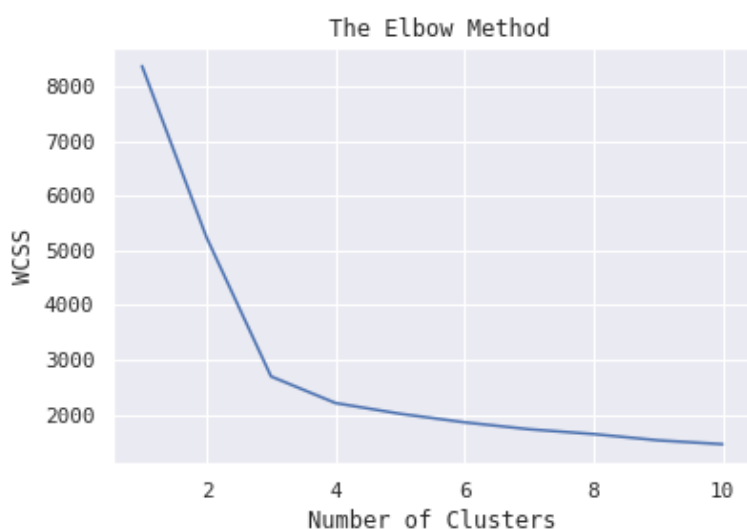
No of clusters: - 3

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Cluster 0

Implied Meaning :- “Object Oriented Programming”

The target word *object* is used in context of OOP. Though in the 1st cluster object is used in the context of an abstract data type, this cluster discerns all the sentences that make use of the word *object* wrt **object oriented programming**.

Example Sentences:-

- 1) In **object-oriented programming languages** like C++, Object Pascal, Java, concept of abstraction become a declarative statement – using keywords `virtual` in C++ `abstract`[11] `interface`[12] in Java.
- 2) hey are typically used as part of an object meta-model. I.e, for each class, defined an instance of the class object in the meta-model is created. Meta-model protocols allow classes to be created and deleted. In this sense, they provide the same functionality as constructors and destructors described above. But in some languages such as the **Common Lisp Object System (CLOS)** the meta-model allows the developer to dynamically alter the object model at run time: e.g., to create new classes, redefine the class hierarchy, modify properties, etc.
- 3) A method in **object-oriented programming (OOP)** is a procedure associated with a message and an object.
- 4) Abstraction in **object-oriented programming**. In object-oriented programming theory, abstraction involves the facility to define objects that represent abstract "actors" that can perform work, report on and change their state, and "communicate" with other objects in the system.

Context Words: -

class, construct, java, application, oriented, computer, method, statement, function, call, static, variables etc.

Threshold Analysis: -

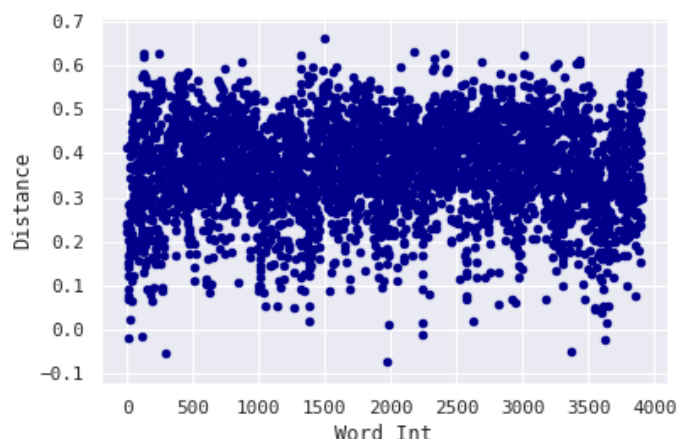
Comprehension Window :- 8

Max Difference : - 0.252

Distance of the farthest relevant word : - 0.489

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Object 224 | 0.365 | 0.689 | 0.365 |
| Object 654 | 0.321 | 0.654 | 0.321 |
| Object 456 | 0.322 | 0.617 | 0.322 |
| Object 214 | 0.362 | 0.559 | 0.3221 |



Plot for label 214 for cluster 0

Cluster 1

Implied Meaning :- “Abstract Data Type”

The target word is used in context of an abstract data type in this cluster. Though the context is OOP but the usage of the word is different. Here, it represents an abstract data type, while in the 1st cluster it depicts or points towards a way of programming.

Example Sentences: -

1) **object Employee** class might contain either directly pointer object Address class, addition instance variables like "first_name" "position". Object composition used represent "has-a" relationships: every employee address, every Employee object access place store Address object either directly embedded within itself, separate location addressed via pointer. Languages support classes almost always support inheritance. This allows classes arranged hierarchy represents "is-a-type-of" relationships.

2) For example, instead inheriting class Person, class Employee could give **Employee object** **internal Person object**, opportunity hide external code even class Person many public attributes methods. Some languages, like Go support inheritance

3) *One of the most important capabilities that a method provides is method overriding - the same name (e.g., area) can be used for multiple different kinds of classes. This allows the sending objects to invoke behaviors and to delegate the implementation of those behaviors to the receiving **object**.*

4) **Consumers object** may consist of various kinds of elements, programs, remote computer systems, computer programmers that they wish to utilize using the object part programs.^ "What Object?". Oracle.com

Context Words:-

attribute, abstractions, classes, instance, container, method, class, artifact, item, body, element, program, java, class, componenet etc.

Threshold Analysis: -

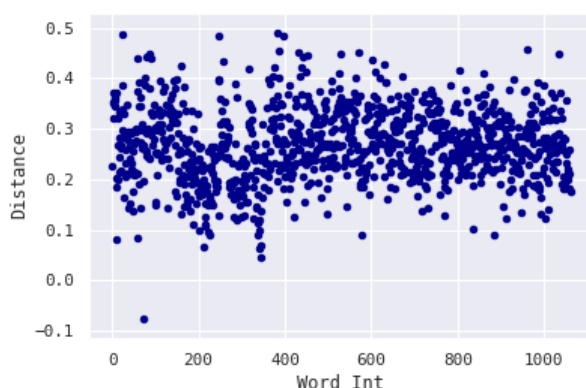
Comprehension Window :- 7

Max Difference : - 0.212

Distance of the farthest relevant word : - 0.475

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Label 254 | 0.315 | 0.654 | 0.32 |
| Label 657 | 0.369 | 0.589 | 0.324 |
| Label 431 | 0.256 | 0.632 | 0.365 |
| Label 270 | 0.227 | 0.599 | 0.321 |



Plot for label 254 of cluster 1

Cluster 2

Implied Meaning :- “Item”, “Artifact”

In this cluster the word *object* has been used as a physical object or an item.

Example Sentences :-

- 1) The deterministic logical cellular automata is necessary as a sufficient condition for the white-box model. The second necessary prerequisite of the white-box model is the presence of a physical **ontology object study**.
- 2) IaaS clouds often offer additional resources such as a virtual-machine disk-image library, raw block storage, file or **object storage**, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles.
- 3) In humans, accomplished motor responses. Spatial planning movement, speech production, complex motor movements aspects action. Consciousness[edit]Main article: ConsciousnessConsciousness awareness whether something **external object** something within oneself.
- 4) Consciousness is the awareness whether something is an **external object** or something within oneself. This helps the mind with having the ability to experience or feel a sense of self.

Context Words: -

item, container, artifact, device, file, pointer, representations, hard, break, solid etc.

Threshold Analysis: -

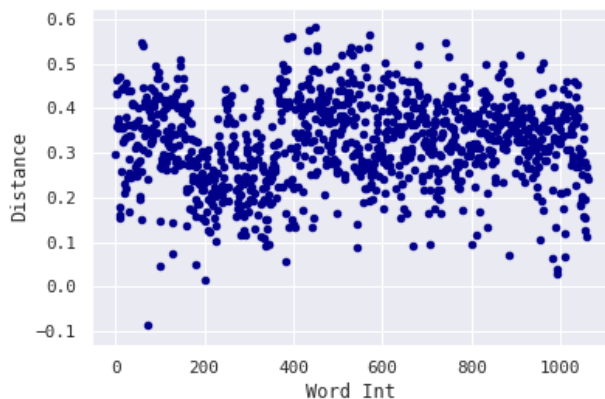
Comprehension Window :- 6

Max Difference : - 0.342

Distance of the farthest relevant word : - 0.484

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Label 254 | 0.345 | 0.627 | 0.31 |
| Label 657 | 0.321 | 0.587 | 0.390 |
| Label 431 | 0.237 | 0.532 | 0.354 |
| Label 270 | 0.201 | 0.699 | 0.313 |



Plot for label 431 of cluster 2

Section 2: - A Few More Examples

Target word: - Operation

No of clusters: - 3

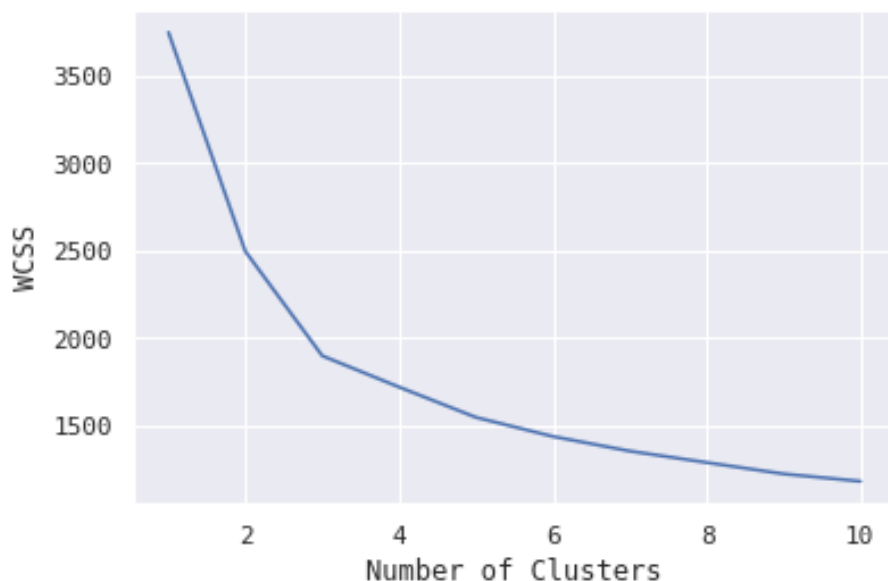
Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES

The Elbow Method



Elbow Method

Cluster 0

Implied Meaning: - “Mathematical Operation”

The context in which the target word has been used in this cluster is related to *mathematics*.

Example Sentences:-

- 1) Two's complement is a **mathematical operation** on binary numbers, and is an example of a radix complement. It is used in computing as a method of signed number representation.
- 2) In computer programming, a **bitwise operation** operates on a bit string, a bit array or a binary numeral (considered as a bit string) at the level of its individual bits. It is a fast and simple action, basic to the higher level arithmetic operations and directly supported by the processor. Most bitwise operations are presented as two-operand instructions where the result replaces one of the input operands.
- 3) The most commonly **studied operations** are binary operations (i.e., operations of arity 2), such as addition and multiplication, and unary operations (i.e., operations of arity 1), such as additive inverse and multiplicative inverse. An operation of arity zero, or nullary operation, is a constant.[2][3] The mixed product is an example of an operation of arity 3, also called ternary operation.
- 4) An operator is similar to an operation in that it refers to the symbol or the process used to **denote the operation**, [12] hence their point of view is different. For instance, one often speaks of "the operation of addition" or "the addition

operation", when focusing on the operands and result, but one switches to "addition operator" (rarely "operator of addition"), when focusing on the process, or from the more symbolic viewpoint, the function $+: X \times X \rightarrow X$.

Context Words: -

combinations, computations, mathematics, functions, operators, prodeure, solving, action, pulse, variable etc.

Threshold Analysis: -

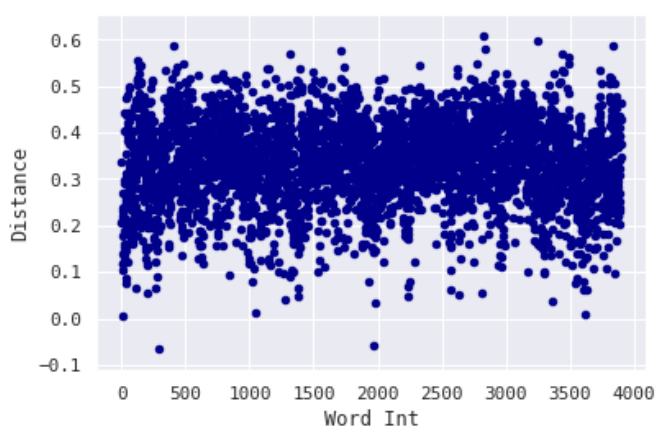
Comprehension Window :- 9

Max Difference : - 0.252

Distance of the farthest relevant word : - 0.4795

Analysis of word vectors

| Word Label | Mean | Max | Median |
|---------------|-------|--------|--------|
| Operation 824 | 0.285 | 0.584 | 0.384 |
| Operation 463 | 0.336 | 0.609 | 0.342 |
| Operation 652 | 0.367 | 0.6616 | 0.3765 |
| Operation 309 | 0.367 | 0.6234 | 0.325 |



Plot for label 463 and 309 of cluster 0

Cluster 1

Implied Meaning :- "Functioning", "Execution".

The target word has been used in the context of functioning and execution in this cluster. The cluster 1 and 2 have the same implication of the word but there is difference in the context in which it is used. In cluster 1 the word *operation* is used with respect to software domain and executable process relating to the software domain.

Example Sentences:-

- 1) To describe algorithms, one usually includes ADT definition **create operation yields** instance ADT, usually axioms equivalent to the result create distinct instance use algorithm. This axiom may strengthen exclude also partial aliasing instances.
- 2) Although average cost per operation is constant and fairly small, the cost of a **single operation** may be quite high.
- 3) The Heapsort algorithm involves preparing the list by first turning it into a max heap. The algorithm then repeatedly swaps the first value of the list with the last value, decreasing the range of values considered in the **heap operation** by one, and sifting the new first value into its position in the heap. This repeats until the range of considered values is one value in length.
- 4) The base class comprises **operations** to compute the sum of the squares between two integers. The subclass re-uses all of the functionality of the base class with the exception of the operation that transforms a number into its square, replacing it with an operation that transforms a number into its square and cube respectively. The subclasses therefore compute the sum of the squares/cubes between two integers.

Context Words: -

Procedure, operate, program, actions, programmer, computer, function, method, compiler, technique, line, code, execution etc.

Threshold Analysis: -

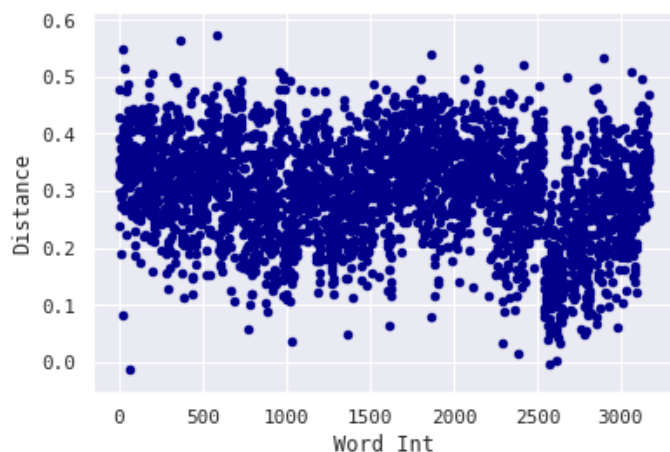
Comprehension Window :- 9

Max Difference : - 0.2827

Distance of the farthest relevant word : - 0.453

Analysis of word vectors

| Word Label | Mean | Max | Median |
|---------------|--------|-------|--------|
| Operation 496 | 0.300 | 0.573 | 0.309 |
| Operation 625 | 0.267 | 0.531 | 0.265 |
| Operation 717 | 0.4020 | 0.657 | 0.209 |
| Operation 790 | 0.3163 | 0.629 | 0.325 |



Plot of label 429 of cluster 1

Cluster 2

Implied Meaning: - “Functioning”, “Execution”, “Procedure”

In this cluster the target word is being used in context of *hardware operations and procedures*.

Example Sentences: -

- 1) **To ensure proper operation of the CPU**, the clock period is longer than the maximum time needed for all signals to propagate (move) through the CPU. In setting the clock period to a value well above the worst-case propagation delay, it is possible to design the entire CPU and the way it moves data around the "edges" of the rising and falling clock signal. This has the advantage of simplifying the CPU significantly, both from a design perspective and a component-count perspective
- 2) The form, design, and implementation of CPUs have changed over time, but their fundamental operation remains almost unchanged. Principal components of a CPU include the arithmetic logic unit (ALU) that performs arithmetic and logic operations, processor registers that supply operands to the ALU and store the results of ALU **operations**, and a control unit that orchestrates the fetching (from memory) and execution of instructions by directing the coordinated operations of the ALU, registers and other components.
- 3) The fundamental **operation** of most CPUs, regardless of the physical form they take, is to execute a sequence of stored instructions that is called a program. The instructions to be executed are kept in some kind of computer memory. Nearly all CPUs follow the fetch, decode and execute steps in their operation, which are collectively known as the instruction cycle.
- 4) The most promising ideas about program-development parallels seem to us to be ones that point to an apparently close analogy between processes within cells, and the low-level **operation** of modern computers.[9] Thus, biological systems are like computational machines that process input information to compute next states, such that biological systems are closer to a computation than classical dynamical system.[10]

Context Words: -

floating, hardware, control, project, interrupt, kernel, load, integrated, punch, execution, usage, performance, inspection, behavior, work

Threshold Analysis: -

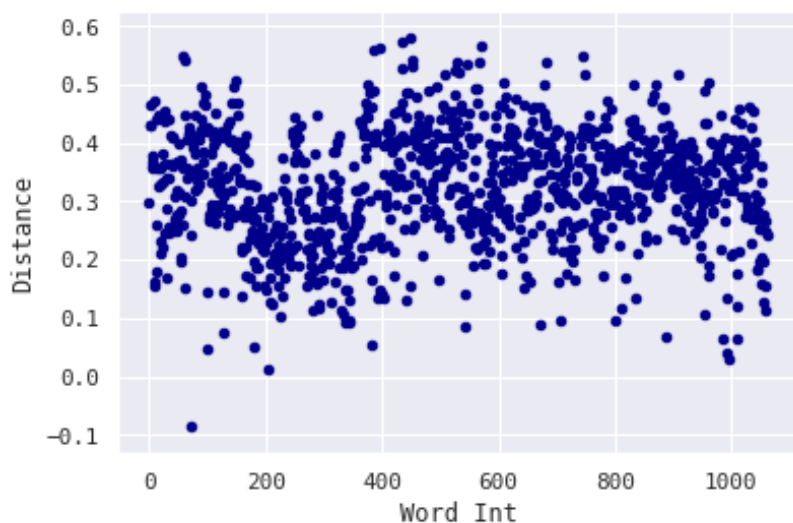
Comprehension Window :- 9

Max Difference : - 0.2145

Distance of the farthest relevant word : - 0.4592

Analysis of word vectors

| Word Label | Mean | Max | Median |
|---------------|--------|-------|--------|
| Operation 622 | 0.242 | 0.479 | 0.236 |
| Operation 541 | 0.261 | 0.490 | 0.267 |
| Operation 48 | 0.301 | 0.647 | 0.306 |
| Operation 326 | 0.3225 | 0.580 | 0.302 |



Plot for label 48 for cluster 2

Target word: - Source

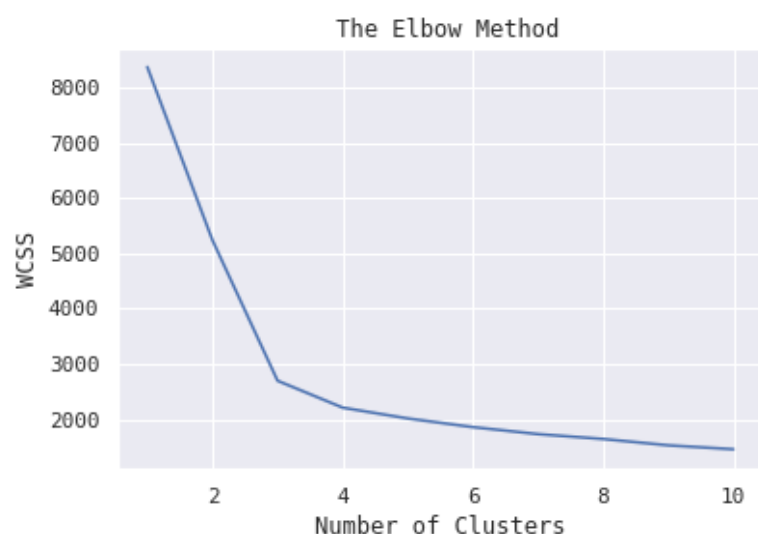
No of clusters: - 3

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Method

Cluster 0

Implied Meaning :- “A place from where data is taken”

In the above cluster the word *source* implies the destination from where data is taken (with respect to computer science and engineering).

Example Sentences:-

- 1) Within each of those standalone components there could be many different **source files**, each containing the program code to handle a part of the problem, with only selected interfaces available to other parts of the program. A sign on program could have source files for each data entry screen and the database interface (which may itself be a standalone third party library or a statically linked set of library routines).
- 2) **Source code** (also referred to as source or code) is the version of software as it is originally written (i.e., typed into a computer) by a human in plain text (i.e., human readable alphanumeric characters).
- 3) Within each of those standalone components there could be many different **source files**, each containing the program code to handle a part of the problem, with only selected interfaces available to other parts of the program. A sign on program could have source files for each data entry screen and the database interface (which may itself be a standalone third party library or a statically linked set of library routines).

Either the database or the payroll application also has to initiate the process of exchanging data with between ship and shore, and that data transfer task will often contain many other components.

4) An API differs from an application binary interface (ABI) in that an API is source code based while an ABI is binary based. For instance, POSIX provides APIs while the Linux Standard Base provides an ABI.

Context Words:

native, destination, root, libraries, environments, path, definition, byte, archive, navigator, run etc.

Threshold Analysis: -

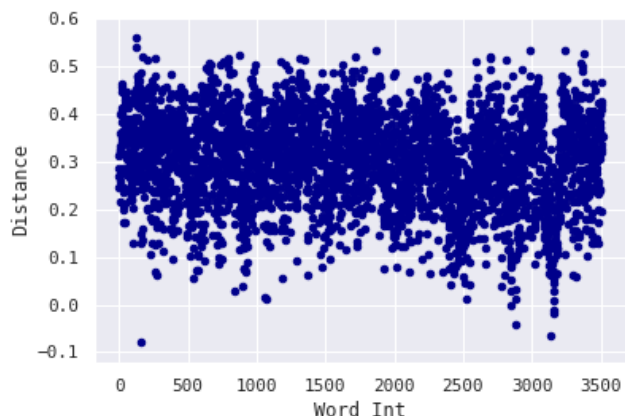
Comprehension Window :- 6

Max Difference : - 0.2455

Distance of the farthest relevant word : - 0.456

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|-------|--------|
| Source 676 | 0.259 | 0.55 | 0.245 |
| Source 678 | 0.3053 | 0.628 | 0.3654 |
| Source 435 | 0.302 | 0.562 | 0.310 |
| Source 172 | 0.231 | 0.227 | 0.227 |



Plot of label 172 for cluster 0

Cluster 1

Implied Meaning : - “Origin Destination”, “Destination of incipience”.

The meaning of the word source in this cluster is in the context of a source *destination* or an *origin destination*.

Example Sentence: -

1) Any data sent across the network requires some time to travel from the **source destination**. Depending on the application, one-way delay round-trip time causes significant impact on the performance.

2) "In order to transmit a digital signal over an analog subscriber line, modulated transmission must be used; that is the electrical signal that represents the binary bit stream of the **source (digital)** output must first be converted to an analog signal that is compatible with a (telephony) speech signal."

3) Many of the protocols in the TCP/IP suite do not provide mechanisms for authenticating **the source** or destination of a message, leaving them vulnerable to spoofing attacks when extra precautions are not taken by applications to verify the identity of the sending or receiving host. IP spoofing and ARP spoofing in particular may be used to leverage man-in-the-middle attacks against hosts on a computer network. Spoofing attacks which take advantage of TCP/IP suite protocols may be mitigated with the use of firewalls capable of deep packet inspection or by taking measures to verify the identity of the sender or recipient of a message.

4) The processing element carries out arithmetic and logical operations, and a sequencing and control unit can change the order of operations in response to stored information. Peripheral devices include input devices (keyboards, mice, joystick, etc.), output devices (monitor screens, printers, etc.), and input/output devices that perform both functions (e.g., the 2000s-era touchscreen). Peripheral devices allow information to be retrieved from an **external source** and they enable the result of operations to be saved and retrieved.

Context Words: -

destination, read, input, sequence, occurrence, compute, published, references, links, protocol, trigger, service

Threshold Analysis: -

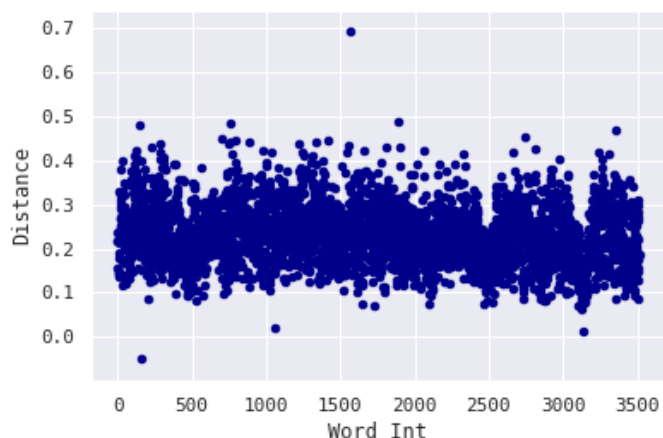
Comprehension Window :- 6

Max Difference : - 0.161

Distance of the farthest relevant word : - 0.4516

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|--------|--------|
| Source 552 | 0.374 | 0.7130 | 0.38 |
| Source 23 | 0.33 | 0.723 | 0.27 |
| Source 22 | 0.33 | 0.723 | 0.27 |
| Source 387 | 0.367 | 0.6876 | 0.345 |



Plot for label 435 for cluster 1

Cluster 2

Implied meaning: - “Free to use”, “Publicly available”, “Freely contributable and mutable”

In this cluster the word *source* is used to denote software for which the original source code is made freely available and may be redistributed and modified.

Example Sentence: -

1) *The open-source model is a decentralized software development model that encourages open collaboration,[10][11] meaning "any system of innovation or production that relies on goal-oriented yet loosely coordinated participants who interact to create a product (or service) of economic value, which they make available to contributors and noncontributors alike."*

2) **Open source code libraries** often contain a testing harness to allow contributors to ensure their changes do not cause regression bugs in the code library

3) **"Open source standardization: rise Linux network era"** PDF. Knowledge, Technology & Policy.

4) *The software can be developed for a variety of purposes, the three most common being to meet specific needs of a specific client/business (the case with custom software), to meet a perceived need of some set of potential users (the case with commercial and **open source software**), or for personal use.*

Context Words:-

Operating, system, release, code, local, open, distribution, written, compilation, git, resources, data, complex, portals, versions, vendor, mutable, script, commit etc.

Threshold Analysis: -

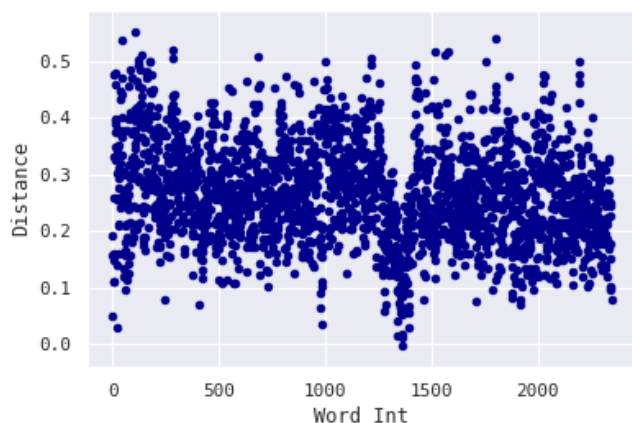
Comprehension Window :- 6

Max Difference : - 0.161

Distance of the farthest relevant word : - 0.4516

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|--------|--------|
| Source 52 | 0.320 | 0.611 | 0.38 |
| Source 9 | 0.219 | 0.533 | 0.198 |
| Source 28 | 0.26 | 0.5411 | 0.256 |
| Source 34 | 0.266 | 0.5500 | 0.262 |



Plot of label 34 for cluster 2

Target word: - State

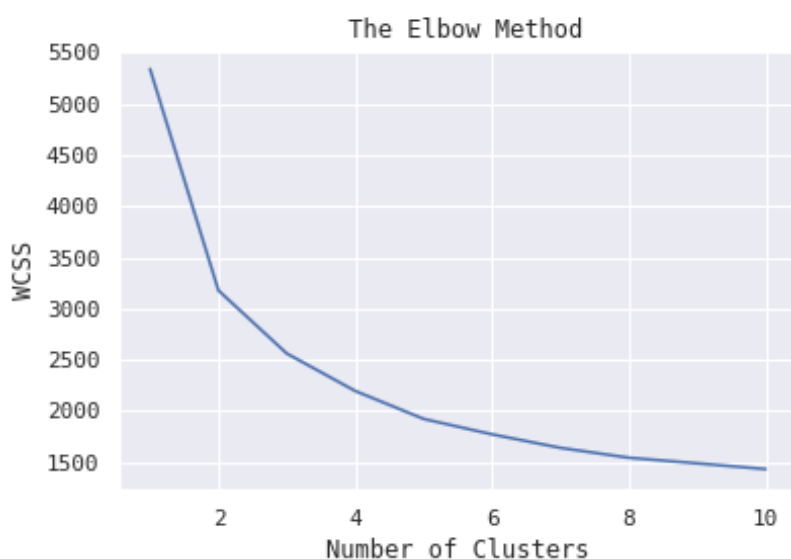
No of clusters: - 2

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Method

Cluster 0

Implied Meaning : - “Phase”, “Situation”, “Condition”

The target word state in this cluster is used to indicate the condition or situation of something. The intended context is discernable. The sentences and the context words collected from the clusters provide an empirical evidence for the same.

Example sentences: -

- 1) The variables are not independent of each other as the state variables are dependent on the decision, input, random, and exogenous variables. Furthermore, the output variables are dependent on the **state of the system** (represented by the state variables).
- 2) The stored contents of these memory elements, at a given point in time, is collectively referred to as the circuit's **state** and contains all the information about the past to which the circuit has access.
- 3) Similarly, a computer program stores data in variables, which represent storage locations in the computer's memory. The contents of these memory locations, at any given point in the program's execution, is called the **program's state**.
- 4) A more specialized definition of **state** is used for computer programs that operate serially or sequentially on streams of data, such as parsers, firewalls, communication protocols and encryption.

Context words: -

phase, moment, define, specific, rate, accessed, rejected, determine, halt etc.

Threshold Analysis: -

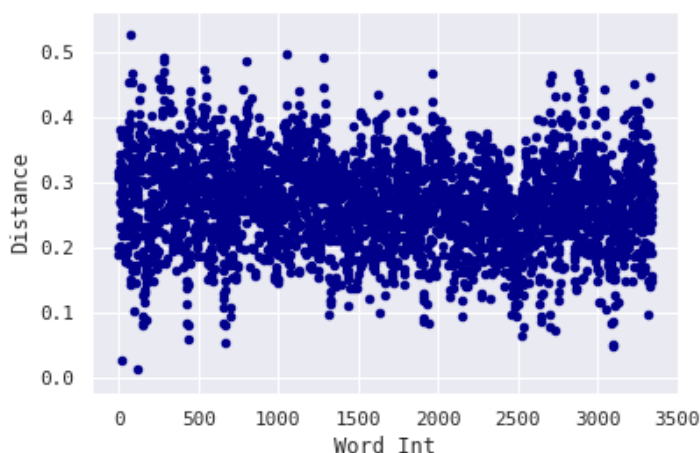
Comprehension Window :- 5

Max Difference : - 0.1287

Distance of the farthest relevant word : - 0.45

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|-------|--------|
| State 48 | 0.2188 | 0.578 | 0.2125 |
| State 80 | 0.1804 | 0.53 | 0.172 |
| State 29 | 0.2244 | 0.590 | 0.2203 |
| State 63 | 0.1988 | 0.550 | 0.187 |



Plot for label 29 for cluster 0

Cluster 1

Implied Meaning: - “Behavioral Software Design Pattern”.

Example Sentence: -

- 1) The **state pattern** can be interpreted as a strategy pattern, which is able to switch a strategy through invocations of methods defined in the pattern's interface.
- 2) Implementing **state-specific** behavior directly within a class is inflexible because it commits the class to a particular behavior and makes it impossible to add a new state or change the behavior of an existing state later independently from (without changing) the class. In this, the pattern describes two solutions

3) The **state pattern** is set to solve two main problems: An object should change its behavior when its internal state changes. State-specific behavior should be defined independently. That is, adding new states should not affect the behavior of existing states.

4) Define separate (state) objects that encapsulate **state-specific** behavior for each state. That is, define an interface (state) for performing state-specific behavior, and define classes that implement the interface for each state.

Context Words: -

structure, object, graph, java, web, structure, design, maps, diagram, plot, class, inherit, pattern etc.

Threshold Analysis: -

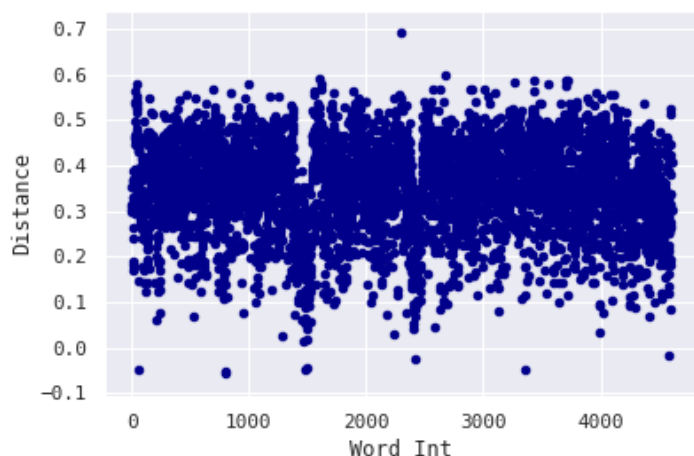
Comprehension Window :- 10

Max Difference : - 0.2004

Distance of the farthest relevant word : - 0.4916

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|--------|--------|
| State 373 | 0.350 | 0.692 | 0.360 |
| State 727 | 0.367 | 0.6917 | 0.375 |
| State 604 | 0.321 | 0.577 | 0.333 |
| State 13 | 0.3259 | 0.6321 | 0.3599 |



Plot for label 13 for cluster 1

Target word: - Table

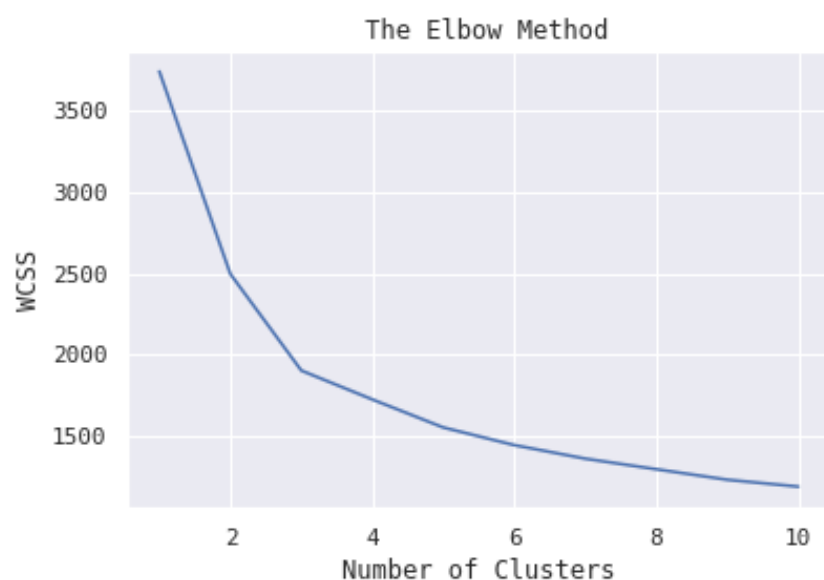
No of clusters: - 3

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Plot

Cluster 0

Implied Meaning : - “Data Structure”

In the above cluster, the word table is used in context of a *hash table* which is a type of data structure.

Example Sentences: -

1) Hash functions are used in conjunction with **Hash table** to store and retrieve data items or data records. The hash function translates the key associated with each datum or record into a hash code which is used to index the hash table.

2) The hash function translates the key associated with each datum or record into a hash code which is used to index the **hash table**. When an item is to be added to **the table**, the hash code may index an empty slot (also called a bucket), in which case the item is added to the table there.

3) A good randomizing function is (barring computational efficiency concerns) generally a good choice as a hash function, but the converse need not be true. **Hash tables** often contain small subsets of valid inputs.

4) In many applications, the range of hash values may be different for each run of the program, or may change along the same run (for instance, when a **hash table** needs to be expanded)

Context Words: - Chart, data type, tree, diagram, columns, stack, pop, push, pointer, range, queue, jump, list etc.

Threshold Analysis: -

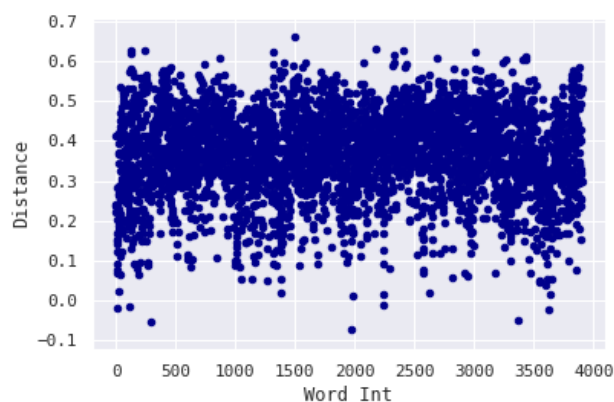
Comprehension Window :- 8

Max Difference : - 0.252

Distance of the farthest relevant word : - 0.489

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Table 224 | 0.365 | 0.689 | 0.365 |
| Table 654 | 0.321 | 0.654 | 0.321 |
| Table 456 | 0.322 | 0.617 | 0.322 |
| Table 214 | 0.362 | 0.559 | 0.3221 |



Plot for label 214 and 456 for cluster 0

Cluster 1

Implied meaning : - “Database”

In this cluster the word *table* indicates or relates to a database or is used in context to a database.

Example Sentence: -

1) One or more columns of **each table** were designated as a primary key by which the rows of the table could be uniquely identified; cross-references between tables always used these primary keys, rather than disk addresses, and queries would join tables based on these key relationships, using a set of operations based on the mathematical system of relational calculus (from which the model takes its name). Splitting the data into a set of normalized tables (or relations) aimed to ensure that each "fact" was only stored once, thus simplifying update operations

2) In this paper, he described a new system for storing and working with large databases. Instead of records being stored in some sort of linked list of free-form records as in CODASYL, Codd's idea was to organize the data as a number of "tables", **each table** being used for a different type of entity. Each table would contain a fixed number of columns containing the attributes of the entity.

3) In the relational approach, the data would be normalized into a user **table**, an address table and a phone number table (for instance). Records would be created in these optional tables only if the address or phone numbers were actually provided.

4) IBM started working on a prototype system loosely based on Codd's concepts as System R in the early 1970s. The first version was ready in 1974/5, and work then started on **multi-table systems** in which the data could be split so that all of the data for a record (some of which is optional) did not have to be stored in a single large "chunk".

Context Words: -

chart, database, list, row, column, structure, cell, display, stack, plot, insert, queue, seed, query, xml, log etc.

Threshold Analysis: -

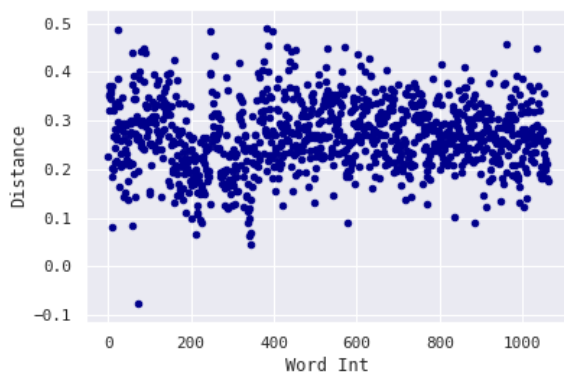
Comprehension Window :- 7

Max Difference : - 0.212

Distance of the farthest relevant word : - 0.475

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Label 254 | 0.315 | 0.654 | 0.32 |
| Label 657 | 0.369 | 0.589 | 0.324 |
| Label 431 | 0.256 | 0.632 | 0.365 |
| Label 270 | 0.227 | 0.599 | 0.321 |



Plot for label 254 and 431 of cluster 1

Cluster 2

Implied Meaning :- “Representation Plot/Type”

In this cluster, table refers to a representation. Used in context of *tabular representation*.

Example Sentence:-

- 1) Any "simple" table can be represented as a "multi-dimensional" table by normalizing the data values into ordered hierarchies. A common example of such a table is a **multiplication table**.
- 2) See **table information** for more information on Windows OS.
- 3) Usage in software specification can encompass ad hoc inclusion of simple **decision tables** in textual documents through the use of tabular specification methodologies, examples of which include SCR[10] and Statestep
- 4) Typical values 1, 4, 8, 16, 24 32.1E 30 4 compression method used. See **next table** list possible values22 34 4 image size. This size raw bitmap data; dummy 0 given BI_RGB bitmaps.26 38 4 horizontal resolution image.

Context Words: -

List, row, column, structure, tree, map, display, chart, representation, section, space, border, interval, matrix, depth, numbers, cells etc.

Threshold Analysis: -

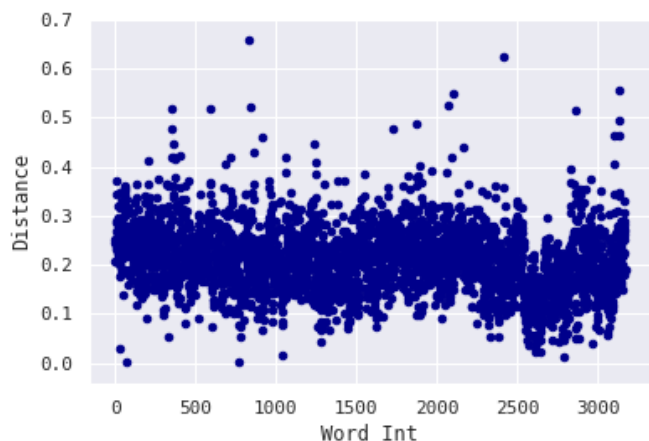
Comprehension Window :- 8

Max Difference : - 0.202

Distance of the farthest relevant word : - 0.452

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|--------|--------|
| Table 245 | 0.325 | 0.632 | 0.312 |
| Table 603 | 0.256 | 0.652 | 0.206 |
| Table 380 | 0.214 | 0.5889 | 0.220 |
| Table 407 | 0.361 | 0.542 | 0.308 |



Plot for label 380 and 407 for cluster 2

Target word: - Window

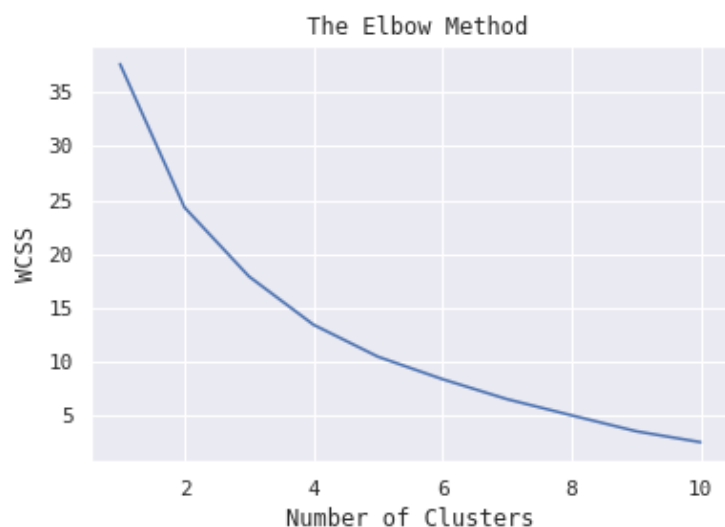
No of clusters: - 2

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Plot

Cluster 0

Implied Meaning : - “Time Frame”, “Interval”

In this cluster the implied sense of the word *window* is with respect to a time frame or a period of respite or interval.

Example Sentences: -

- 1) An uninterruptible power supply like UPS is used to give a computer a brief **window of time** to move information from primary volatile storage to non-volatile storage until the batteries are exhausted.
- 2) In some applications, such as substring search, one can compute a hash function h for every k -character substring of a given n -character string by advancing a **window** of width k characters along the string; where k is a fixed integer, and n is greater than k .
- 3) These include: exponential backoff protocols 802.11's CSMA/CA original Ethernet, **window reduction** TCP, fair queueing devices routers. Another method avoid negative effects network congestion implementing priority schemes, packets transmitted higher.

Context words:-

disclosure, time, move, width, short, tracks, environment, shell, queue, advancing, frame etc.

Threshold Analysis: -

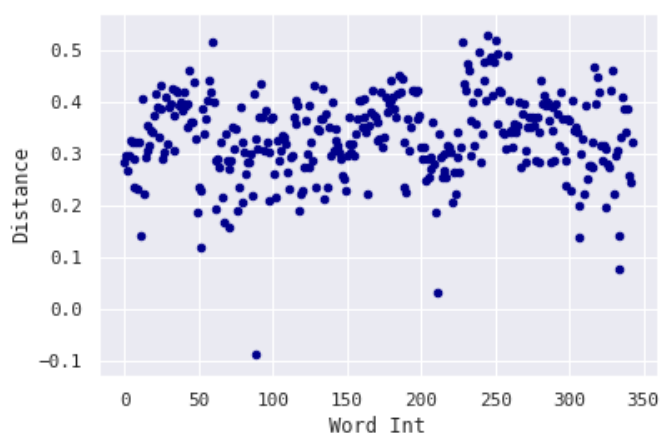
Comprehension Window :- 5

Max Difference : - 0.14437

Distance of the farthest relevant word : - 0.467

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|-------|--------|
| Window 37 | 0.326 | 0.611 | 0.3299 |
| Window 38 | 0.3169 | 0.612 | 0.321 |
| Window 45 | 0.338 | 0.541 | 0.334 |
| Window 25 | 0.333 | 0.530 | 0.338 |



Plot for label 18 and 25 of cluster 0

Cluster 1

Implied Meaning :- “An Opening”

Though the occurrence of the context is rare in this context, one event was found where this context of the target word was used. The clustering algorithm clustered it accordingly which is an extremely remarkable outcome, and exhibits the tangibility of the approach.

Example Sentence: -

*1) Once programmed, an EPROM can be erased by exposing it to strong ultraviolet light source (such as from a mercury-vapor lamp). EPROMs are easily recognizable by the transparent fused quartz (or on later models resin) **window** on the top of the package, through which the silicon chip is visible, and which permits exposure to ultraviolet light during erasing*

Context Words:-

glass, water, clean, wall, building, upgrade, little, light, erased, area, blinds etc.

Threshold Analysis: -

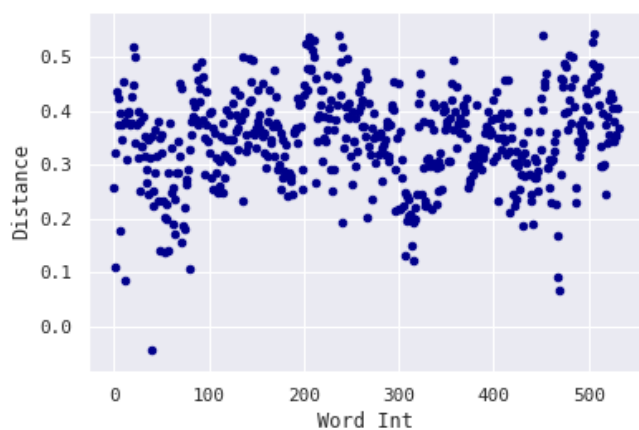
Comprehension Window :- 6

Max Difference : 0.384

Distance of the farthest relevant word : - 0.421

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| Window 224 | 0.346 | 0.544 | 0.354 |
| Window 654 | 0.321 | 0.654 | 0.321 |
| Window 456 | 0.322 | 0.617 | 0.322 |
| Window 214 | 0.362 | 0.559 | 0.3221 |



Plot for Label 18 of cluster 1

Target word: - Machine

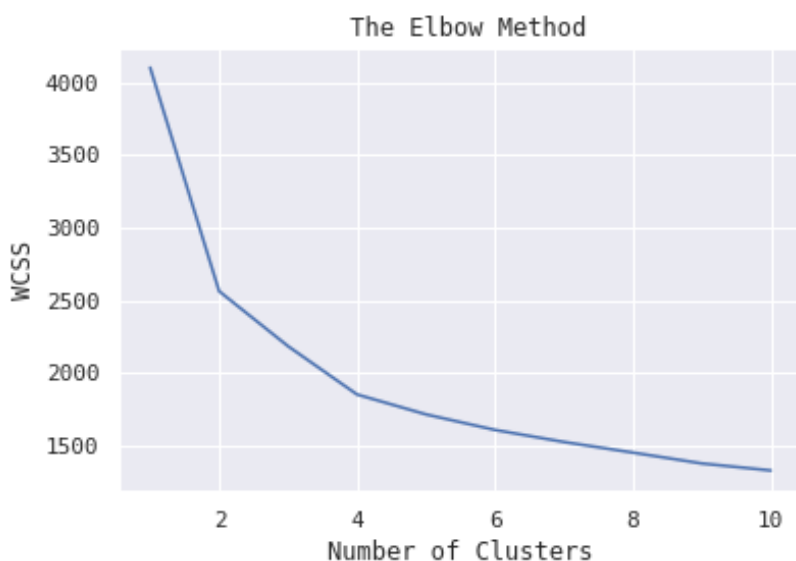
No of clusters: - 4

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES



Elbow Method

Cluster 0

Implied Meaning: - “Machine Learning”, “Artificial Intelligence”

The results obtained for the target word *machine* are extremely confounding. In this cluster the word machine has been used in context to **artificial intelligence and machine learning**.

Example Sentences: -

- 1) Some example classes of search algorithms, sorting algorithms, merge algorithms, numerical algorithms, graph algorithms, string algorithms, computational geometric algorithms, combinatorial algorithms, medical algorithms, **machine learning**, cryptography, data compression algorithms parsing techniques.
- 2) These sub-fields are based on technical considerations, such as particular goals (e.g. "robotics" or "**machine learning**"), the use of particular tools ("logic" or artificial neural networks), or deep philosophical differences.
- 3) Faster computers, algorithmic improvements, and access to large amounts of data enabled advances in **machine learning** and perception; data-hungry deep learning methods started to dominate accuracy benchmarks around 2012.[57] The Kinect, which provides a 3D body-motion interface for the Xbox 360 and the Xbox One, uses algorithms that emerged from lengthy AI research[58] as do intelligent personal assistants in smartphones

4) Some straightforward applications of natural language processing include information retrieval, text mining, question answering[119] **machine translation**. [120] Many current approaches use word co-occurrence frequencies construct syntactic representations text.

Context Words :-

computation, intelligence, robots, human, robotic, engine, reality, computers, automation, automated, neural, image, processing, vision, pattern, natural, text etc.

Threshold Analysis: -

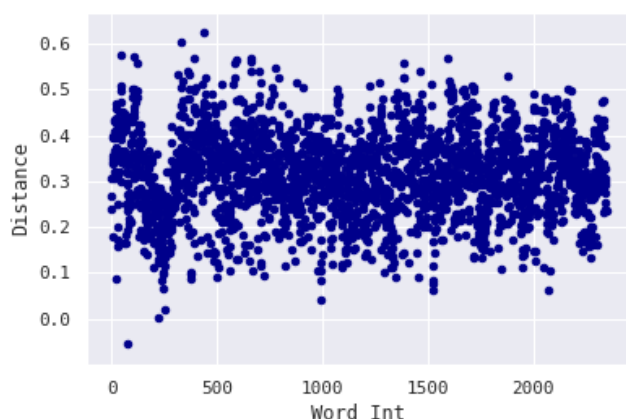
Comprehension Window :- 7

Max Difference : - 0.249

Distance of the farthest relevant word : - 0.45

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-------------|--------|--------|--------|
| Machine 155 | 0.311 | 0.6327 | 0.3172 |
| Machine 277 | 0.287 | 0.5781 | 0.287 |
| Machine 610 | 0.299 | 0.570 | 0.3020 |
| Machine 193 | 0.2752 | 0.6244 | 0.2757 |



Plot for label 155 of cluster 0

Cluster 1

Implied Meaning :- “Mechanical Apparatus”, “Mechanical Device”.

The target word **machine** is used in context of a mechanical apparatus.

Example Sentences: -

- 1) What is meant by "behavior" varies by author, with the two main types of formal specifications for behavior being axiomatic (algebraic) specification and an abstract model; these correspond to axiomatic semantics and operational semantics of an **abstract machine**, respectively. Some authors also include the computational complexity ("cost"), both in terms of time (for computing operations) and space (for representing values).
- 2) the establishment and use of sound engineering principles in order to economically obtain software that is reliable and works efficiently on a real **machine**.
- 3) When the first digital computers appeared in the early 1940s, the instructions to make them operate were wired into the **machine**. Practitioners quickly realized that this design was not flexible and came up with the "stored program architecture" or von Neumann architecture.
- 4) A typical abstract **machine consists** definition terms input, output, set allowable operations used turn former latter. The best-known example **Turing machine**.

Context Words: -

machinery, engine, computer, hardware, wire, representation, compiler, mechanical, motion, vector, system, processing, input, registers etc.

Threshold Analysis: -

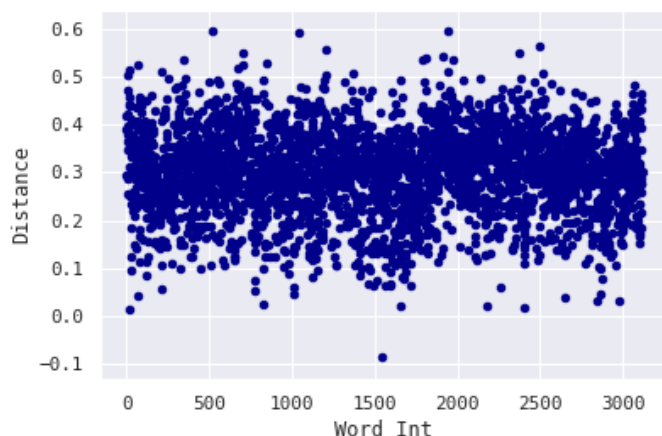
Comprehension Window :- 11

Max Difference : - 0.196

Distance of the farthest relevant word : - 0.4661

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-------------|--------|--------|--------|
| Machine 300 | 0.358 | 0.690 | 0.366 |
| Machine 412 | 0.313 | 0.568 | 0.3249 |
| Machine 107 | 0.3477 | 0.6109 | 0.358 |
| Machine 26 | 0.3014 | 0.594 | 0.307 |



Plot for label 26 of cluster 1.

Cluster 2

Implied Meaning :- “Computer”, “Intelligent Machine”

The target word is used in the software domain typically representing a computer or a piece of machinery that is **intelligent and automated in nature**. The word is related to **automation and robotics** and basically it is the context domain of this cluster.

Example Sentences: -

- 1) The authorities feed all this data into **an artificial-intelligence machine** that rates people's loyalty to the Communist Party in order to control every aspect of their lives.
- 2) As the **machine** receives a new input, it looks at the state and picks a new spot based on the information on what to do when it receives that input at that state. When there are no more inputs, the automaton stops and the space it is on when it completes determines whether the automaton accepts or rejects that particular set of inputs.
- 3) The German encryption machine, Enigma, was first attacked with the help of the electro-mechanical bombs which were often run by women.[30][31] To crack the more sophisticated German Lorenz SZ 40/42 machine, used for high-level Army communications, Max Newman and his colleagues commissioned Flowers to build the Colossus.
- 4) he first modern analog computer was a **tide-predicting machine**, invented by Sir William Thomson 1872.

Context words : - Device, computer, equipment, programming, intelligent, automation, system, process, data, electronic, functional, computing, learning, robot, robotic etc.

Threshold Analysis: -

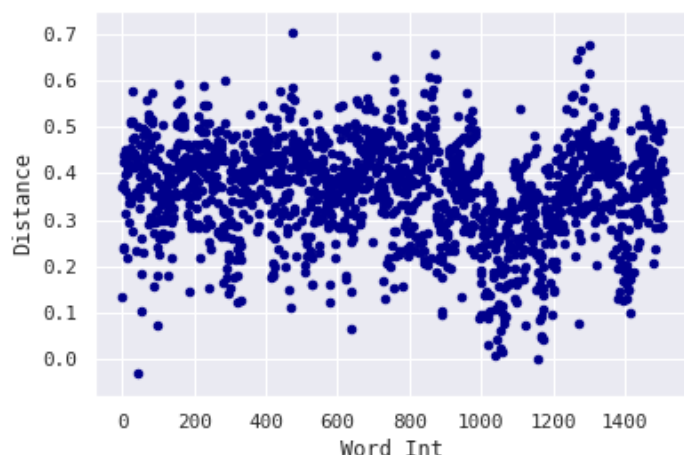
Comprehension Window :- 9

Max Difference : - 0.458

Distance of the farthest relevant word : - 0.4612

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-------------|--------|--------|--------|
| Machine 362 | 0.3588 | 0.677 | 0.3588 |
| Machine 612 | 0.3316 | 0.6340 | 0.3400 |
| Machine 360 | 0.3641 | 0.7026 | 0.378 |
| Machine 352 | 0.3541 | 0.5987 | 0.3625 |



Plot for label 290 of cluster 2

Cluster 3

Implied Meaning : - “Programming Language”, “Machine Assembly Language”.

The target word *machine* is used in context of a programming language. The belaboring occurrence of the word in this cluster can be found in the context of ***assembly language programming***.

Example Sentences: -

1) Computer languages can be processed with a computer. An example of this abstraction process is the generational development of programming languages from the **machine language** to the assembly language and the high-level language. Each stage can be used as a stepping stone for the next stage.

2) Each level uses a system of expression involving a unique set of objects and compositions that apply only to a particular domain. [14] Each relatively abstract, "higher" level builds on a relatively concrete, "lower" level, which tends to provide an increasingly "granular" representation. For example, gates build on electronic circuits, binary on gates, **machine language** on binary, programming language on machine language, applications and operating systems on programming languages

3) For example, gates build on electronic circuits, binary on gates, machine language on binary, programming language on **machine language**, applications and operating systems on programming languages. Each level is embodied, but not determined, by the level beneath it, making it a language of description that is somewhat self-contained.

4) A much more human friendly rendition of **machine language**, called assembly language, uses mnemonic codes to refer to machine code instructions, rather than using the instructions' numeric values directly, and uses symbolic names to refer to storage locations and sometimes registers

Context Words: -

assembly, machine, language, operating, development, software, microsoft, log, registers, processors, development, primitive, write, electronic etc.

Threshold Analysis: -

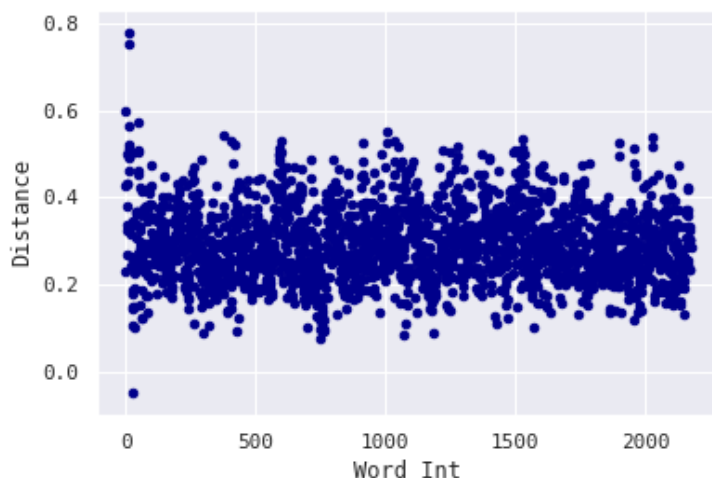
Comprehension Window :- 10

Max Difference : -0.319

Distance of the farthest relevant word : - 0.459

Analysis of word vectors

| Word Label | Mean | Max | Median |
|-------------|---------|--------|--------|
| Machine 728 | 0.277 | 0.545 | 0.281 |
| Machine 5 | 0.2938 | 0.777 | 0.2868 |
| Machine 373 | 0.31450 | 0.531 | 0.3198 |
| Machine 490 | 0.3209 | 0.5816 | 0.384 |



Plot for label 373 of cluster 3

Section 3: - Examples of Overlapping-Clusters

Target word: - Interface

No of clusters: - 4

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES (2 out of 4) (Cluster 0 and Cluster 1)

NO(2 out of 4) (Cluster 2 and Cluster 3)

Recommended Clusters After Analysis :-2

The clusters formed for the target word *interface* are quite ambiguous. The ambiguity stems in the 3rd and the 4th cluster, as it gets difficult to discern them from the 1st and the 2nd cluster respectively. The 1st cluster implies a similar sense of the target word when compared to the remaining clusters. But, though the implied meaning of the word remains the same, this approach sheds light on the context of the domain it has been used in. Onto closer examination and analysis, we can group the word into clusters, based on its usage in the *software and hardware* domain.

The Elbow Method



Elbow Method

Cluster 0,2 (As they have the same sense of the word *Interface*)

Implied Meaning :- “Interaction”, “Platform / Periphery where interaction takes place ”

In all the recommended clusters the implied context of the word is same, and it becomes difficult to discern them from one rather. This seems to be an issue of the clustering algorithm rather than the vector embedder (BERT). The predictions obtained from the *elbow method* seem to be misleading in this case.

But the approach yields remarkable results in delineating the hardware and the software contexts of the word.

The first cluster has grouped the usage of the word *interface* in the **software** domain. On analyzing the context words, predicted by the distance similarity function, this hypothesis is empirically established.

Example Sentences: -

1) Early multi-user DBMS typically only allowed for the application to reside on the same computer with access via terminals or terminal emulation software. The client-server architecture was a development where the application resided on a client desktop and the database on a server allowing the processing to be distributed. This evolved into a multitier architecture incorporating application servers and web servers with the end **user interface** via a web browser with the database only directly connected to the adjacent tier

2) This **interface** has been popularized with the Lego Mindstorms system and is being actively perused by a number of companies wishing to capitalize on the power of custom browsers like those found at Mozilla. KTechlab supports flowcode and is a popular open-source IDE and Simulator for developing software for microcontrollers

3) One option for **interface** testing is to keep a separate log file of data items being passed, often with a timestamp logged to allow analysis of thousands of cases of data passed between units for days or weeks. Tests can include checking the handling of some extreme data values while other interface variables are passed as normal values.

4) A **brain-computer interface** (BCI), sometimes called a neural control interface (NCI), mind-machine interface (MMI), direct neural interface (DNI), or brain-machine interface (BMI), is a direct communication pathway between an enhanced

or wired brain and an external device. BCIs are often directed at researching, mapping, assisting, augmenting, or repairing human cognitive or sensory-motor functions.

Context Words :- *interaction, library, api, border, connection, abstractions, drivers, parameters, model, program, application*

Threshold Analysis: -

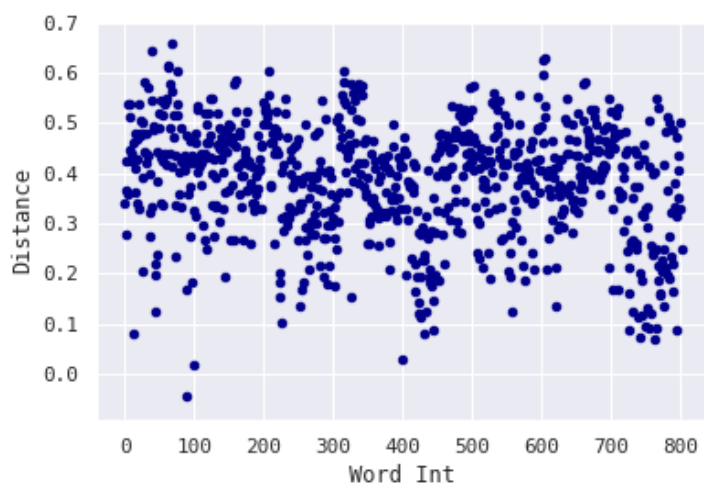
Comprehension Window :- 11

Max Difference : - 0.1915

Distance of the farthest relevant word : - 0.4585

Analysis of word vectors

| Word Label | Mean | Max | Median |
|---------------|--------|--------|---------|
| Interface 210 | 0.384 | 0.657 | 0.403 |
| Interface 614 | 0.3536 | 0.618 | 0.3666 |
| Interface 383 | 0.3425 | 0.5855 | 0.3568 |
| Interface 266 | 0.3287 | 0.5863 | 0.33206 |



Plot for label 210 of cluster 0

Cluster 1,3

As stated above, the implied sense of the target word remains same, though the domain in which it is used changes. Cluster number 1 and 4 determine the use of the word *interface* in context of the hardware domain. Empirical evidence can be obtained by observing the predicted context words.

Example Sentences:-

1) Microcoding has been popular in application-specific processors such as network processors, microcontrollers, digital signal processors, channel controllers, disk controllers, network **interface controllers**, graphics processing units, and in other hardware

2) CPU, core memory and external **bus interface** of a DEC PDP-8/I, made of medium-scale integrated circuits.

3) Most modern CPU designs are at least somewhat superscalar; and nearly all general purpose CPUs designed in the last decade are superscalar. In later years some of the emphasis in designing high-ILP computers has been moved out of the CPU's hardware **interface**, or instruction set architecture (ISA). The strategy of the very long instruction word (VLIW) causes some ILP to become implied directly by the software, reducing the amount of work the CPU must perform to boost ILP and thereby reducing the design's complexity.

4) Some components, such as **bus interface** and cache, may be shared between cores. Because the cores are physically close to each other, they can communicate with each other much faster than separate (off-chip) processors in a multiprocessor system, which improves overall system performance.

Context Words :-

connection, interaction, connector, link, wiring, switch, frame, binding, driver, bus, mask, mechanism, hardware, instructions etc.

Threshold Analysis: -

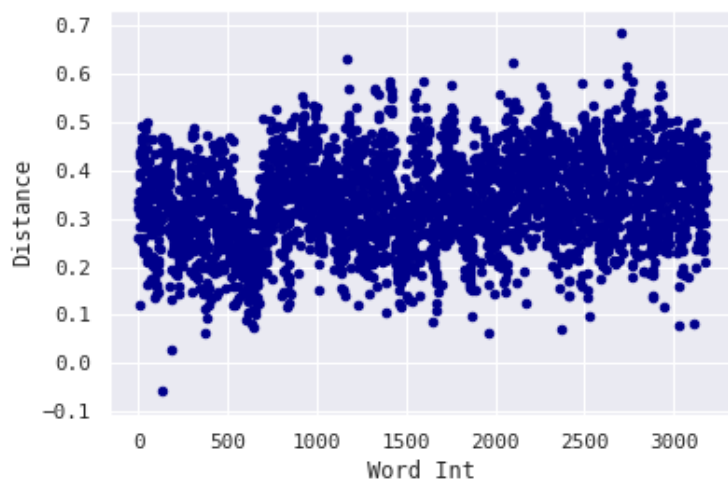
Comprehension Window :- 8

Max Difference : - 0.1907

Distance of the farthest relevant word : - 0.4947

Analysis of word vectors

| Word Label | Mean | Max | Median |
|---------------|--------|---------|--------|
| Interface 111 | 0.336 | 0.684 | 0.3375 |
| Interface 253 | 0.3261 | 0.611 | 0.3341 |
| Interface 631 | 0.333 | 0.6031 | 0.342 |
| Interface 405 | 0.3271 | 0.58066 | 0.3398 |



Plot for label III of cluster 1

Target word: - System

No of clusters: - 3

Deterministic Method :- Elbow Method

Clustering Algorithm: - K-Means Clustering

Do the number of clusters make sense?

YES (2 out of 3)(Cluster 0 &1)

NO (1 out of 3) (Cluster 2)

Recommended no of clusters after analysis? 2

The Elbow Method



Elbow Method

Cluster 0 and Cluster 2

Implied Meaning: - “Structure”, “Organization”

The implied context of the target word is with respect to a set of things working together as parts of a mechanism or an interconnecting network; a complex whole.

Example Sentences :-

- 1) To the ancients, the parallel postulate seemed less obvious than the others. They aspired to create a **system** of absolutely certain propositions, and to them it seemed as if the parallel line postulate required proof from simpler statements
- 2) The variables represent some properties of the **system**, for example, measured system outputs often in the form of signals, timing data, counters, event occurrence (yes/no). The actual model is the set of functions that describe the relations between the different variables.
- 3) The question of whether the model describes well the properties of the **system** between data points is called interpolation, and the same question for events or data points outside the observed data is called extrapolation.
- 4) Mathematical models are usually composed of relationships and variables. Relationships can be described by operators, such as algebraic operators, functions, differential operators, etc. Variables are abstractions of **system** parameters of

interest, that can be quantified. Several classification criteria can be used for mathematical models according to their structure

Context words: -

organization, environment, facility, collection, agent, network, architecture, components, service, etc.

Threshold Analysis: -

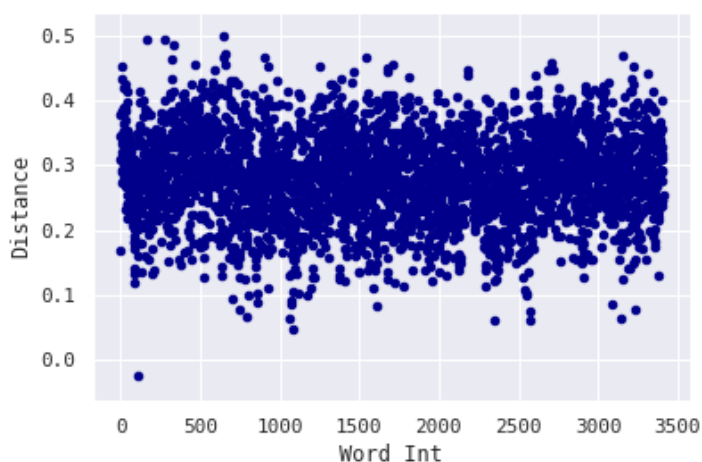
Comprehension Window :- 0.13

Max Difference :- -0.202

Distance of the farthest relevant word :- 0.4579

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|--------|-------|--------|
| System 651 | 0.35 | 0.695 | 0.365 |
| System 652 | 0.0283 | 0.599 | 0.254 |
| System 658 | 0.273 | 0.524 | 0.27 |
| System 720 | 0.33 | 0.634 | 0.357 |



Plot for label 652 of cluster 0

Cluster 1

Implied Meaning : - “an organized scheme or method.”

The implied meaning of the target word differs from the initial cluster and the context it is used in is totally different. This cluster displays the use of the word system in context to an organization or a scheme or a stratagem.

Example Sentences:-

- 1) The contemporary national **legal systems** are generally based on one of four basic systems: civil law, common law, statutory law, religious law or combinations of these.
- 2) Components of the life-**support system** are life-critical, and are designed and constructed using safety engineering techniques.
- 3) Nearly all organisms have some kind of immune system. Bacteria have a rudimentary **immune system** in the form of enzymes that protect against virus infections. Other basic immune mechanisms evolved in ancient plants and animals and remain in their modern descendants.
- 4) He defined a **social system** as only a segment (or a "subsystem") of what he called action theory.[4] Parsons organized social systems in terms of action units, where one action executed by an individual is one unit.

Context Words: -

Unit, Structure, agent, base, agency, interchanges, functions, codes, conditions, methodology, action, entire, society, resources, comprehensive, operated etc.

Threshold Analysis: -

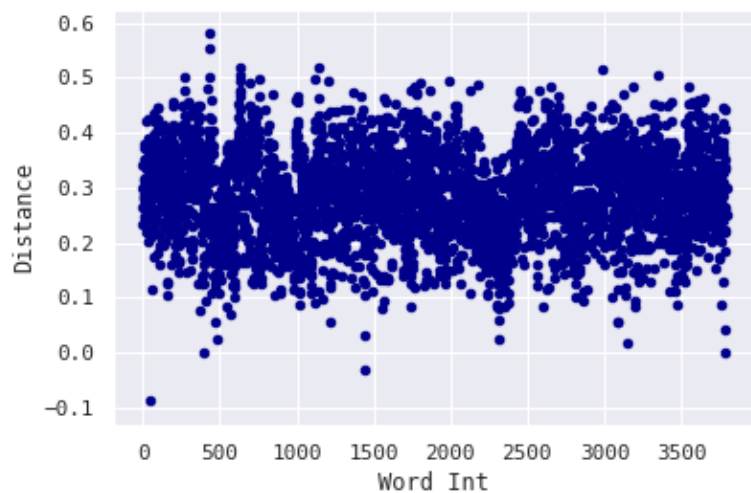
Comprehension Window :- 5

Max Difference : - 0.2361

Distance of the farthest relevant word : - 04579

Analysis of word vectors

| Word Label | Mean | Max | Median |
|------------|-------|-------|--------|
| System 49 | 0.329 | 0.689 | 0.339 |
| System 57 | 0.264 | 0.564 | 0.265 |
| System 79 | 0.256 | 0.579 | 0.287 |
| System 91 | 0.289 | 0.58 | 0.298 |



Plot for labels 57 of cluster 1

END