ကန့်သတ်

ရုရှားဖက်ဒရေးရှင်းနိုင်ငံ၊ စိန့်ပီတာစဘတ်မြို့

Saint Petersburg National Research University of Information Technologies, Mechanics and Optic (ITMO) တက္ကသိုလ်တွင် ဆောင်ရွက်ခဲ့သော မဟာဘွဲ့သင်တန်းတက်ရောက်စဉ် သုတေသနပြု ဆောင်ရွက်ခြင်းဆိုင်ရာ မဟာဘွဲ့ယူကျမ်း

(အကျဉ်းချုပ်)

ပြန်တမ်းဝင်အမှတ်၊ ကြည်း ၇၄၈၇၀

အဆင့်၊ ဗိုလ်ကြီး

အမည်၊ ကျော်ဇေယျနိုင်

ရာထူး၊ တပ်စုမှူး

တပ်၊ ခမရ (၂၀၆)၊ သထုံ (ကြာပန်း)

ရက်စွဲ၊ ၂၀၂၅ ခုနှစ်၊ ဩဂုတ်၊ ( ) ရက်။

နေရာ၊ စတခ(၃)၊ရွာတော်၊နေပြည်တော်။

ကန့်သတ်

နိဒါန်း

၁။ ကျွန်တော် ပြန်တမ်းဝင်အမှတ် ကြည်း ၇၄၈၇၀ ဗိုလ်ကြီး ကျော်ဇေယျနိုင်သည် ရုရှားနိုင်ငံ၊ စိန့်ပီတာစဘတ်မြို့၊ Saint Petersburg National Research University of Information Technologies, Mechanics and Optics တက္ကသိုလ်တွင် အသုံးချကွန်ပျူတာသိပ္ပံ မာစတာဘွဲ့ (Master of Applied Computer Sciences) သင်တန်းကို တက်ရောက်ခဲ့ပါသည်။ ကျွန်တော်ရွေးချယ်ထားခဲ့သည့် စာတမ်းခေါင်းစဉ် " Development of a web application for plagiarism analysis of reports on educational projects " ဖြင့် ဘွဲ့ယူကျမ်း ပြုစုခဲ့ပါသည်။

ရည်ရွယ်ချက်

၂။ မဟာဘွဲ့ကျမ်း၏ ရည်ရွယ်ချက်မှာ Bachelor, Master, PhD ကျောင်းသားများ၏ သုတေသနစာတမ်းများတွင် စာကူးယူမှု (Plagiarism) ရာခိုင်နှုန်းကို တိကျစွာစစ်ဆေးကာ၊ မူရင်းအကြောင်းအရာများ ပိုမိုဖန်တီးနိုင်စေရန်အတွက် နည်းပညာအထောက်အပံ့ပေးသည့် စနစ်တစ်ခု ဖွံ့ဖြိုးရန်ဖြစ်ပါသည်။ ထို့အပြင် အဖွဲ့လိုက်လုပ်ဆောင်ရသော ပရောဂျက်များတွင် ကျောင်းသားများ ရေးသားသော ပရောဂျက် အစီရင်ခံစာများကို အချင်းချင်းခိုးကူးယူမှုနည်းပါးစေရန် အစီရင်ခံစာများအတွင်းရှိ တူညီမှုရှိသော စာကြောင်းများ ကို Highlightပြုလုပ်ပြီး တူညီမှုရာခိုင်နှုန်းကို ဖော်ပြပေးပြီး analysis ပြုလုပ်နိုင်ရန် ရိုးရှင်းသော interface ကို ဖန်တီးထားခြင်းဖြစ်ပါသည်။ စာတမ်း၏ အဓိကရည်ရွယ်ချက်များမှာ အောက်ပါအတိုင်းဖြစ်ပါသည်-

(က) **စာကူးယူမှု ရာခိုင်နှုန်းကို တိကျစွာ ခွဲခြမ်းစိတ်ဖြာခြင်း** - TF-IDF၊ Sentence-BERT (S-BERT) နှင့် Hybrid Algorithm (0.3×TF-IDF + 0.7×S-BERT) တို့ကို အသုံးပြု၍ ဝေါဟာရတူညီမှုနှင့် အဓိပ္ပာယ်တူညီမှု နှစ်မျိုးလုံးကို စစ်ဆေးခြင်းနှင့် ကျောင်းသားများ၏ အစီရင်ခံစာများကို နှိုင်းယှဉ်ကာ တူညီသောအပိုင်းများကို Highlight ပြုလုပ်ပေးခြင်း။

(ခ) **မူရင်းစာတမ်းများ ဖန်တီးနိုင်စေရန် အားပေးခြင်း** – စာကူးယူမှုများကို အလိုအလျောက်ဖော်ထုတ်ခြင်းဖြင့် ကျောင်းသားများ သူတို့ကိုယ်ပိုင်အတွေးအခေါ်များ ဖွံ့ဖြိုးစေရန်နှင့် သုတေသနစာတမ်းများ၏ အရည်အသွေးနှင့် မူရင်းပါဝင်မှု မြင့်မားလာစေရန်။

(ဂ) **ပညာရေးဆိုင်ရာ တင်းကျပ်မှုများကို ထောက်ကူခြင်း** - တက္ကသိုလ်များ၊ ဆရာများအနေဖြင့် ကျောင်းသားများ၏ အလုပ်များကို လွယ်ကူစွာ စိစစ်နိုင်စေရန်နှင့် ပညာရေး စံနှုန်းများ မြှင့်တင်ရာတွင် အထောက်အကူဖြစ်စေရန်။

သင်ကြားခဲ့သော ဘာသာရပ်များ

၃။ သင်ကြားခဲ့သော ဘာသာရပ်များမှာ အောက်ပါအတိုင်းဖြစ်ပါသည်။ -

စဉ်	ဘာသာရပ်များ	သင်ကြားချိန်စုစုပေါင်း(နာရီ)	ရလာဒ်
၁	Methodology for organizing design and development of information systems	၁၀၈	Good
၂	Network architecture and cloud technologies	၁၀၈	Passed
၃	International Research Management Essentials	၁၀၈	Passed
၄	Foreign Language	၁၀၈	Passed
၅	Big Data: Storage Technologies and Elements of Statistics	၁၀၈	Good
၆	High Tech Business Creation: check-list for entrepreneurs	၁၀၈	Good
၇	Cloud data storage and processing	၁၀၈	Excellent
၈	Network architecture and cloud technologies (Course Project)	၁၀၈	Excellent
၉	System analysis and modeling of information processes and systems	၁၀၈	Excellent
၁၀	Automatic text processing	၁၀၈	Passed

၁၁	Mathematical data processing methods	၁၀၈	Passed
၁၂	Foreign Language	၁၀၈	Passed
၁၃	Translation research methodology	၁၀၈	Passed
၁၄	Introduction to Machine Learning (tools) and Applied Artificial Intelligence in Science and Business	၁၀၈	Excellent
၁၅	Applied data analysis packages	၂၁၆	Good
၁၆	Deep learning (Course Project)	၂၁၆	Excellent
၁၇	Decision support systems development	၁၀၈	Excellent
၁၈	Mobile application programming technologies	၁၀၈	Good
၁၉	Research work	၉၇၂	Excellent
၂၀	Industrial and technological practice (project-technological)	၂၁၆	Excellent
၂၁	Production, pre-diploma	၂၁၆	Excellent
၂၂	Graduate qualification work (master's dissertation)	၅၄၀	Excellent

လေ့လာသင်ယူခြင်း

၄။ ဤသုတေသနလုပ်ငန်းကို ရည်ရွယ်ချက်အောင်မြင်စေရန် အောက်ပါလုပ်ငန်းများကို အဆင့်ဆင့် ဆောင်ရွက်ခဲ့ပါသည်။ ပထမဦးစွာ အသုံးပြုသူများ စနစ်အတွင်းသို့ လုံခြုံစွာ ဝင်ရောက်နိုင်ရန် Authentication လုပ်ငန်းစဉ်ကို သတ်မှတ်ပေးထားပါသည်။ ထို့နောက် PDF, Docx နှင့် Txt ဖိုင်အမျိုးအစားများမှ စာသားအချက်အလက်များကို ထိရောက်စွာ စွဲထုတ်နိုင်ရန် နည်းပညာများ ဖော်ဆောင်ထားခဲ့ပါသည်။ စာသားများကို စနစ်တကျ စီမံရန် text preprocessing

အဆင့်များဖြစ်သည့် Tokenization, Lemmatization နှင့် stop word များဖယ်ထုတ်ခြင်းစသည့် လုပ်ငန်းများကို အတိအကျ ဆောင်ရွက်ထားပါသည်။ ထို့နောက် Sentence Transformer နည်းပညာများကို အသုံးပြု၍ စာသားများ၏ အဓိပ္ပာယ်ဆိုင်ရာ ဆက်စပ်မှုများကို တွက်ချက်သုံးသပ်ပါသည်။ ထို့နောက် စစ်ဆေးရရှိလာသော ရလဒ်များကို Data Visualization နည်းလမ်းများဖြင့် ရှင်းလင်းစွာ ပြသပေးနိုင်ရန် စီစဉ်ထားပါသည်။ နောက်ဆုံးအဆင့်အနေဖြင့် နှိုင်းယှဉ် လေ့လာပြီး အစီရင်ခံစာများကို စနစ်တကျထုတ်ပေးခြင်းနှင့် Database အတွင်း သိမ်းဆည်းထားနိုင်ရန် ဆောင်ရွက်ထားပါသည်။

အသုံးပြုခဲ့သည့်သီအိုရီနှင့် ဖြေရှင်းဆောင်ရွက်မှု

၅။ ဤသုတေသနတွင် Bachelor, Master, PhD ကျောင်းသားများ၏ သုတေသန အစီရင်ခံစာများအကြား တူညီသည့်အကြောင်းအရာများကို ထိရောက်စွာ ရှာဖွေနိုင်ရန် အဓိကသီအိုရီ နည်းလမ်းနှစ်ရပ်ကို ပေါင်းစပ်အသုံးပြုထားသည်။

(က) ပထမနည်းလမ်းမှာ Lexical method (TF-IDF + cosine similarity) ဖြစ်ပြီး ၎င်းတွင် TF (Term Frequency) ဖြင့် စာကြောင်းထဲတွင် ဝေါဟာရတစ်ခု ဖြစ်ပေါ်မှုအကြိမ်အရေအတွက်ကို တိုင်းတာခြင်း၊ IDF (Inverse Document Frequency) ဖြင့် ဝေါဟာရတစ်ခု၏ ထူးခြားမှုအဆင့်ကို အကဲဖြတ်ခြင်းတို့ ပါဝင်သည်။ ဤနည်းလမ်းသည် ထူးခြားနေသော စကားလုံးများကို ပိုမိုအရေးပေးပြီး အများသုံးစကားလုံးများ၏ အရေးပါမှုကို လျော့နည်းစေရန် ချိန်ညှိပေးနိုင်သည်။ စာသားများကို ဗက်တာပုံစံပြောင်းလဲပြီး ကျိုးမြေတြီနည်းဖြင့် နှိုင်းယှဉ်ခြင်းဖြင့် ရလဒ်များရရှိနိုင်သည်။

(ခ) ဒုတိယနည်းလမ်းမှာ အဓိပ္ပာယ်ဆိုင်ရာခွဲခြမ်းစိတ်ဖြာမှု (SentenceTransformer) ဖြစ်ပြီး ခေတ်မီ AI နည်းပညာဖြစ်သော Sentence-BERT ကို အခြေခံထားသည်။ ဤနည်းလမ်းသည် စာကြောင်းများကို ရှုထောင့်ပေါင်းစုံမှ အဓိပ္ပာယ်ခွဲခြမ်းစိတ်ဖြာပေးနိုင်ပြီး စကားလုံးများ မတူညီသော်လည်း အဓိပ္ပာယ်တူညီသော ဝါကျများကို ထိရောက်စွာ ရှာဖွေနိုင်စွမ်းရှိပါသည်။ ထို့အပြင် စာကြောင်းများ၏ အကြောင်းအရာနှင့် ရည်ရွယ်ချက်ကို နားလည်နိုင်စွမ်းလည်း ရှိပါသည်။

စာတမ်းအကျဉ်းချုပ်

၆။ ဤသုတေသနလုပ်ငန်းကို အောက်ပါအဆင့်များဖြင့် စနစ်တကျ ဆောင်ရွက်ခဲ့ပါသည်။ ပထမအဆင့်တွင် ပုံစံအမျိုးမျိုးမှ စာသားများကို ထုတ်ယူခြင်းလုပ်ငန်းကို ဆောင်ရွက်ပါသည်။ ဤအဆင့်တွင် PDF၊ DOCX၊ TXT အစရှိသော ဖိုင်ဖော်မတ်အမျိုးမျိုးမှ စာသားအချက်အလက်များကို ထုတ်ယူခြင်း၊ ခေါင်းစဉ်စာမျက်နှာများ၊ အကိုးအကားများ၊ နောက်ဆက်တွဲများကဲ့သို့သော သုတေသနစာတမ်းများ၏ ဘုံတူညီသောအပိုင်းများကို ဖယ်ရှားခြင်း၊ စာသားများကို စံသတ်မှတ်ရန် တိုကင်ယူခြင်း(Tokenization)နှင့် (Lemmatization) ခြင်းတို့ကို ဆောင်ရွက်ပါသည်။ ထို့နောက် စီမံဆောင်ရွက်ပြီးသော စာရွက်စာတမ်းများ၏ ဖွဲ့စည်းပုံဒေတာဘေ့စ်ကို ဖန်တီးပါသည်။

ဒုတိယအဆင့်တွင် Hybrid Similarity Algorithm ကို အသုံးပြုပါသည်။ ဤအဆင့်တွင် အခြေခံကူးယူခြင်းကို ခွဲခြားသတ်မှတ်ရန် TF-IDF ခွဲခြမ်းစိတ်ဖြာမှုကို အသုံးပြုပြီး အဓိပ္ပါယ်ညီမျှမှုကို ရှာဖွေရန် SentenceTransformer ကို အသုံးပြုပါသည်။ ရလဒ်များကို အလေးချိန် ဖော်မြူလာ (0.3 × TF-IDF + 0.7 × SentenceTransformer) ဖြင့် ပေါင်းစပ်ကာ စာရွက်စာတမ်းများအကြား အလုံးစုံတူညီမှုရာခိုင်နှုန်းကို တွက်ချက်ပါသည်။

တတိယအဆင့်တွင် ရလဒ်များကို ပုံဖော်ခြင်းလုပ်ငန်းကို ဆောင်ရွက်ပါသည်။ ဤအဆင့်တွင် ဒက်ရှ်ဘုတ်တစ်ခုပေါ်တွင် အသေးစိတ်အချက်အလက်များကို တင်ပြခြင်း၊ တူညီမှုအဆင့် သတ်မှတ်ချက်အလိုက် အရောင်အမျိုးမျိုးဖြင့် ကိုက်ညီသောအပိုင်းများကို highlight ပြခြင်း၊ အစီရင်ခံစာ၏ စာကြောင်းတူညီမှုကို အပြည့်အစုံ၊ တစ်စိတ်တစ်ပိုင်း၊ မရှိဟူ၍ အသေးစိတ်ပြသခြင်းနှင့် ခွဲခြမ်းစိတ်ဖြာခြင်းအတွက် ပြီးပြည့်စုံသော အစီရင်ခံစာများကို ထုတ်ပေးခြင်းတို့ ပါဝင်ပါသည်။

စာတမ်းမှ ရရှိလာသည့် ရလဒ်များ

၇။ မိမိ၏ သုတေသနစာတမ်းမှ အောက်ပါအကျိုးရလဒ်များကို ရရှိခဲ့ပါသည်။

(က) လက်တွေ့အကောင်အထည်ဖော်မှုဆိုင်ရာ ရလဒ်များအနေဖြင့် "Russian Reading Comprehension with Commonsense Reasoning" dataset ကို အသုံးပြု၍ စနစ်၏တိကျမှုကို စမ်းသပ်ခဲ့ရာတွင် အောင်မြင်စွာ စစ်ဆေးနိုင်ခဲ့သည်။ စာတမ်းကြီးကြပ်သူ (supervisor) မှ

အသုံးပြုခွင့်ပေးထားသော bachelor ကျောင်းသားများ၏ အစီရင်ခံစာ (၁၀၀) ကျော်ကို မှန်ကန်တိကျစွာ စစ်ဆေးနိုင်ခဲ့ပြီး ဤလုပ်ငန်းစဉ်၏ ထိရောက်မှုကို သက်သေပြနိုင်ခဲ့ပါသည်။

(ခ) လွယ်ကူစွာ အသုံးပြုနိုင်သော User Interface တစ်ခုကို အောင်မြင်စွာ ဖန်တီးခဲ့ပြီး ဖြစ်ပါသည်။ ဤ interface သည် အသုံးပြုသူများအား ရိုးရှင်းသော လမ်းညွှန်မှုများဖြင့် စနစ်ကို အဆင်ပြေစွာ အသုံးပြုနိုင်စေရန် ရည်ရွယ်ပြီး သုတေသနရလဒ်များကို ထိရောက်စွာ ကြည့်ရှုနိုင်စေမည် ဖြစ်ပါသည်။

အသုံးပြုနိုင်မည့်နယ်ပယ်များ

၈။ ဤသုတေသနလုပ်ငန်းသည် ပညာရေးနယ်ပယ်တွင် လက်တွေ့အသုံးချနိုင်ပါသည်။ ပထမဦးစွာ ဤစနစ်ကို တက္ကသိုလ်များနှင့် သုတေသနဌာနများတွင် ကျောင်းသားများ၏ ဘွဲ့လွန်စာတမ်းများ၊ သုတေသနစာတမ်းများတွင် မူရင်းအကြောင်းအရာများ ပါဝင်မှုကို စစ်ဆေးရန်နှင့် ပါမောက္ခများနှင့် စာတမ်းကြီးကြပ်သူများအတွက် ကျောင်းသားစာတမ်းများကို အရည်အသွေး စစ်ဆေးရန် အထောက်အကူပြု နည်းပညာအဖြစ်အသုံးပြုနိုင်ပါသည်။ ဆရာ၊ဆရာမများအနေဖြင့် ကျောင်းသားများ၏ သုတေသနစာမ်းများကို လွယ်ကူမြန်ဆန်စွာ စိစစ်နိုင်မည်ဖြစ်ပြီး မူရင်းအကြောင်း အရာများ ဖန်တီးရေးသားနိုင်စွမ်းကို မြှင့်တင်ပေးနိုင်မည်ဖြစ်ပါသည်။ ဒုတိယအနေဖြင့် ဤသုတေသန လုပ်ငန်းသည် ဘာသာစကားအမျိုးမျိုးကို ပံ့ပိုးနိုင်သည့် နည်းပညာတစ်ရပ်ဖြစ်သောကြောင့် ဘာသာစကားအမျိုးမျိုးဖြင့် ရေးသားထားသော စာတမ်းများကို စစ်ဆေးရာတွင်လည်း အသုံးဝင်မည်ဖြစ်သည်။ ဤစနစ်သည် အွန်လိုင်းပလပ်ဖောင်းတစ်ခုအဖြစ် လွယ်ကူစွာဝင်ရောက် အသုံးပြုနိုင်သည့် စနစ်တစ်ခုဖြစ်သောကြောင့် အင်တာနက်အခြေပြု ပညာရေး ဝန်ဆောင်မှုများတွင် လွယ်ကူစွာ ပေါင်းစပ်အသုံးပြုနိုင်မည်ဖြစ်ပါသည်။

လိုအပ်ချက်နှင့်အကြံပြုတင်ပြချက်များ

၉။ ဤ Plagiarism Analysis Web Application စနစ်အတွက် အောက်ပါအကြံပြု တင်ပြချက်များကို ဆောင်ရွက်သင့်ပါသည်။ ပထမဦးစွာ နည်းပညာပိုင်းဆိုင်ရာအနေဖြင့် Hybrid Algorithm၏ တိကျမှုနှုန်းကို ပိုမိုမြှင့်တင်နိုင်ရန် အဆင့်မြင့် Machine Learning မော်ဒယ်များကို အဆက်မပြတ် မွမ်းမံသုတေသနပြုလုပ်ရန်လိုအပ်ပါသည်။ စနစ်၏ လုံခြုံရေးနှင့် စွမ်းဆောင်ရည်ကို

မြင့်တင်ရန် Cloud Computing အခြေပြု ဝန်ဆောင်မှုများနှင့် ပေါင်းစပ်အသုံးပြုသင့်ပါသည်။ နည်းပညာဖွံ့ဖြိုးတိုးတက်မှုနှင့်အညီ Deep Learning နည်းပညာများကို အဆက်မပြတ် မွမ်းမံသင်ယူကာ Paraphrasing ဖမ်းယူနိုင်မည့်စနစ်များ ထည့်သွင်းသင့်ပါသည်။ မြန်မာဘာသာ စကားအတွက် NLP နည်းပညာကို ပိုမိုဖွံ့ဖြိုးအောင် ဆောင်ရွက်သင့်ပါသည်။ ဒုတိယအနေဖြင့် တက္ကသိုလ်များ၊ သုတေသနဌာနများနှင့် ပူးပေါင်း၍ ဤစနစ်ကို Plagiarism စစ်ဆေးရေး စံနှုန်းတစ်ခုအဖြစ် အတည်ပြုနိုင်ရန် ကြိုးပမ်းသင့်ပါသည်။

ရက်စွဲ၊ ၂၀၂၅ ခုနှစ် ဩဂုတ်၊ ( ) ရက်။

နေရာ၊ စတခ(၃)၊ရွာတော်၊နေပြည်တော်။

ကြည်း ၇၄၈၇၀၊ ဗိုလ်ကြီး ကျော်ဇေယျနိုင်

အမှတ်(-)တပ်ရင်း၊နေပြည်တော်။